View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Archive

THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Mutation frequencies in a birth-death branching process

David Cheek

Doctor of Philosophy
University of Edinburgh
2019

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*David Cheek*)

Acknowledgements

I would like to thank my supervisor Tibor Antal, and fellow PhD students Stefano Avanzini and Michael Nicholson. Thank you for your support, comradery, and comedy.

Abstract

A growing population of cells accumulates genetic mutations. We study stochastic models of this process. Cells divide and die as a branching process, and a cell's genetic information is a sequence of nucleotides which mutates randomly at division. Motivated by biologically realistic parameters, we consider that few cells grow to many cells and mutation rates are small, proving approximations in this limit. In particular we are interested in mutation frequencies and their dependency structure along the genetic sequence; the relevance of the evolutionary tree and selection are discussed. Amongst other results, we recover a power-law distribution for mutation frequencies which is consistent with previously published cancer genetic data.

Lay summary

A tumour can grow from one cell to billions of cells. Over billions of cell divisions, errors in DNA replication accumulate. So a tumour is genetically diverse. This diversity is important for at least two reasons. First, it is a key factor in resistance to treatment and disease recurrence. Second, genetic data can provide a window into the past evolutionary trajectory of a tumour. On both counts, mathematics has helped our collective understanding. Mathematical models offer precise, quantitative descriptions which, when combined with data, illuminate otherwise obscure genetic processes. This thesis is not so concerned with data. Rather we study some of the most simple and fundamental probabilistic illustrations of a growing cell population (not limited to cancer). Our broad intention is to explore the quantitative relationship between a cell population's growth and its genetic information.

Contents

Chapter 1

Introduction

1.1 Background and overview

With advances in DNA sequencing technology, vast quantities of cancer genetic data have been made available in recent years. From this data a prominent message is delivered, repeatedly, in the biological literature: cancers exhibit genetic diversity. That diversity exists between different cancers of different individuals is perhaps unsurprising. But more ominously, diversity is a hallmark feature of any single tumour. It is said that the evolution of a tumour is akin to Darwinian evolution, that cells are mutating to provide a diverse population on which selective pressures can act. For example, when a cancer treatment arrives, some cells are resistant to the treatment and continue to proliferate. For this reason and others, understanding the genetic evolution of tumours is a tremendous, ongoing research effort.

One especially common form of data gives mutation frequencies in an individual tumour at a single snapshot in time. The data takes the form $(x_i)_{i\in\mathcal{S}}$, where $i \in \mathcal{S}$ indexes sites on the genome, and x_i is the frequency of mutations at site i (a mutation is a difference from some reference genome). An example is presented in Figure 1.1. What, if anything, can this data teach us about the evolution of the tumour? This question is one motivation for our work. However it should be emphasised that statistical analysis of mutation frequency data is not our game. Rather we wish to explore the question from a mathematical viewpoint. The idea is that simple mathematical models can offer predictions for data. Moreover models can give insight into the relevance of, and the relationships between, features of the evolutionary process. In particular, what can be said about the interplay between the population dynamics (the cell divisions and deaths governing the growth trajectory) and the genetic information?

Our biological motivation is in fact broader than cancer. Any growing population of cells sees cell divisions and at cell divisions there are errors in DNA replication. Thus diversity is generated. This is true for a population of bacterial cells. And it is with bacterial cells that our mathematical story begins.

Luria and Delbrück, in their famous work of 1943 [32], considered a growing population of bacterial cells which is sensitive to attack by a lethal virus. The bacteria may mutate to become resistant to the virus. Their mathematical model

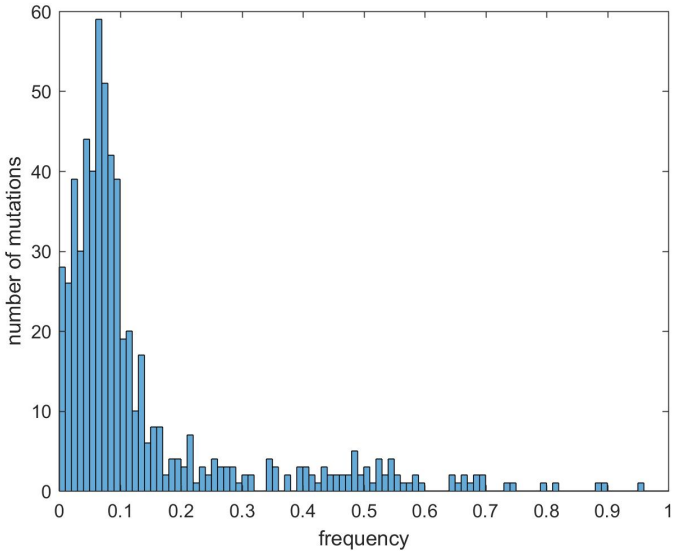


Figure 1.1: A histogram of mutation frequencies from a lung adenocarcinoma [17].

said that cells grow deterministically and exponentially, and that mutation times follow a Poisson process. They compared the model's predictions to experimental data, leading to biological insight and a nobel prize. Lea and Coulson adapted the model, saying that mutant cells grow as a branching process [31]. They obtained a obtained a probability distribution for the number of mutant cells, commonly known as the Luria-Delbrück distribution. The distribution has seen empirical evidence and become a standard tool for the estimation of mutation rates in bacteria [37]. In subsequent decades, the model and adaptations have been the subject of much research (see [41] for a review). In particular, we note that Kendall described cells growing as birth-death branching processes [25].

Kendall's two-type branching process, often referred to as the stochastic Luria-Delbrück model, has been foundational in the mathematical understanding of cancer evolution. The model and various extensions have been used to study drug resistance [20, 28, 16, 6], driver mutations [13, 12], and metastasis [33, 18, 34, 9], for example. As introduced by Kendall, wildtype (type A) and mutant (type B) cells are assumed to divide, die, and mutate independently of each other, according to

$$\begin{cases} A \to AA, & \text{rate } \alpha_A; \\ A \to \emptyset, & \text{rate } \beta_A; \\ A \to AB, & \text{rate } \nu; \\ B \to BB, & \text{rate } \alpha_B; \\ B \to \emptyset, & \text{rate } \beta_B. \end{cases}$$

This model is so simple and fundamental that its scope of applications encompasses more than just genetic change in cells. Anyhow, the total number of

mutants is of key interest. In recent years, [2, 20, 29, 26, 27, 3, 19] derived exact and approximate distributions for the number of mutants at fixed times and population sizes.

In Chapter 2, we offer a rigorous account of Kendall's model. A variety of limit results are proven, for large population sizes, large times, and small mutation rates. The limits approximate the number of mutants, mutation times, and clone sizes (a clone is a subpopulation of mutant cells initiated by a mutation). Both previously known and new results are presented. To conclude the chapter we present an extension of the model. Genetic information is extended from binary (A or B) to a sequence. Each entry of the sequence is binary, depicting whether a position on the genome is mutated. Many results can immediately be extended to this setting. Some results are consistent with previously published cancer genetic data.

In Chapter 3, we further extend the model, but narrowing our focus to a specific process of genetic change. A cell's genetic information is now seen as a finite sequence of the nucleotides A, C, G, and T. Each entry of the sequence can mutate independently at cell division. The mutation model is a little developed from Chapter 2 in that backward mutations and double mutations (a cell divides to give two mutated daughter cells) are allowed. Although we are unaware of other works which state a model of this exact form, we claim no originality. The model is a combination of two truly classic parts: (1) a branching process; (2) the Jukes-Cantor model (or at least a relative of).

This model stands in contrast to recent works which predict mutation frequencies in cancer [7, 10, 35, 39]. They do not describe the genome as a finite sequence of nucelotides. Instead they employ the infinite-sites assumption (ISA). The ISA states that every new mutation must occur at a novel genetic site, which allows a reduced picture of genetic information. The ISA is of immense value. It gifts analytic and computational tractability. However we believe that a 'finite-sites' model (as described in the previous paragraph) is worthy of exploration too, for several reasons:

- Recent statistical analysis refutes the ISA in human cancers [30].
- A finite-sites model can clearly depict the relationship between a cell's genetic sequence and its division and death rates.
- A finite-sites model can clearly depict site-specific mutation rates.
- A finite-sites model is mathematically rich, inviting questions such as: What is the dependency structure between sites? What happens when the number of sites tends to infinity?

Chapter 3 sees an exploration of the finite-sites model. Our primary work is to prove approximations for mutation frequency distributions, and to highlight the relevance of the underlying evolutionary tree. To conclude, for an example application, we estimate mutation rates in a lung adenocarcinoma.

1.2 Preliminaries

Write $\mathbb{N} = \{1, 2, ..\}$ for the positive integers, and $\mathbb{N}_0 = \{0, 1, 2, ..\}$ for the non-negative integers.

Any continuous-time Markov process shall be regarded as a random element of the space of cadlag functions. Denote the space of cadlag functions from $I \subset [0, \infty]$ to R as $\mathbb{D}(I, R)$, which is equipped with the standard Skorokhod topology. Typically R will be $\mathbb{R}^n$. Let's note a property of the space $\mathbb{D}(I, R)$ which is described in Billingsley's excellent book [5].

Lemma 1.2.1. *Suppose that*

1. $(t_n)_{n\in\mathbb{N}} \subset I$ *with* $\lim_{n\to\infty} t_n = t \in I$,
2. $(f_n)_{n\in\mathbb{N}} \subset \mathbb{D}(I, R)$ *with* $\lim_{n\to\infty} f_n = f \in \mathbb{D}(I, R)$, *and*
3. f *is continuous at* t.

Then $\lim_{n\to\infty} f_n(t_n) = f(t)$.

A *birth-death branching process with birth and death rates* α *and* β is a continuous-time Markov process on $\mathbb{N}_0$ with transition rates

$$i \mapsto \begin{cases} i+1, & \text{rate } i\alpha; \\ i-1, & \text{rate } i\beta. \end{cases}$$

Let's note some basic properties of a birth-death branching process $(X(t))_{t\geq 0}$ with birth and death rates α and β. The following lemmas are in [4] or [11] for example.

Lemma 1.2.2. *If* $X(0)$ *is fixed, the generating function of* $X(t)$ *is*

$$\mathbb{E}[z^{X(t)}] = \begin{cases} \left(\frac{\beta(z-1)-e^{-(\alpha-\beta)t}(\alpha z-\beta)}{\alpha(z-1)-e^{-(\alpha-\beta)t}(\alpha z-\beta)}\right)^{X(0)}, & \alpha \neq \beta; \\ \left(\frac{tz-t-z}{tz-t-1}\right)^{X(0)}, & \alpha = \beta. \end{cases}$$

Lemma 1.2.2 can be proved by solving the Kolmogorov equations.

Lemma 1.2.3. *Suppose that* $\alpha > \beta$. *Then*

$$\lim_{t\to\infty} e^{-(\alpha-\beta)t} X(t) = W$$

almost surely, where

$$W \stackrel{d}{=} \sum_{i=1}^{X(0)} \chi_i \psi_i,$$

and the $\chi_i \sim$ *Bernoulli*$(1 - \beta/\alpha)$ *and the* $\psi_i \sim$ *Exponential*$(1 - \beta/\alpha)$ *are independent.*

The convergence in Lemma 1.2.3 can be proved by showing that $e^{-(\alpha-\beta)t}X(t)$ is a martingale, with bounded moments. The limiting distribution can be found by taking the limit of the generating function in Lemma 1.2.2.

Lemma 1.2.4. *The events* $\{\exists t \geq 0, X(t) = 0\}$ *and* $\{W = 0\}$ *are almost surely the same.*

Lemma 1.2.4 can be proved by noting firstly that $\{\exists t \geq 0, X(t) = 0\} \subset \{W = 0\}$, and secondly that $\mathbb{P}[\exists t \geq 0, X(t) = 0] = \mathbb{P}[W = 0]$ (which follows from Lemmas 1.2.2 and 1.2.3).

Chapter 2

Kendall's two-type branching process

This chapter sees an investigation of Kendall's two-type branching process, also known as the stochastic Luria-Delbrück model. We study the convergence of mutation times, the number of mutants, and clone sizes in various limits. In Section 2.1, the model is introduced. In Section 2.2, large-time and large-population almost sure convergence results are presented. In Section 2.3, the simultaneous large-population and small-mutation limit is presented. In Section 2.4, the simultaneous large-time and small-mutation limit is presented. In Section 2.5, exact results are presented. In Section 2.6, the model is extended to multiple genetic sites, and the site frequency spectrum is studied. Section 2.6 can be thought of as a warm-up for Chapter 3. In Section 2.7, proofs are given. In Section 2.8, the results are discussed in relation to other works and data.

2.1 Model

Kendall's model describes the growth trajectory of two types of cells, wildtype and mutant. The wildtype cells grow as a birth-death branching process with birth and death rates α_A and β_A. Write $A(t)$ for the number of wildtype cells at time t, and say that $A(0) \in \mathbb{N}$ is fixed.

Wildtype cells mutate at rate $\nu > 0$. So mutation times follow a Cox process with intensity $(\nu A(t))_{t \geq 0}$ (a Cox process is just a Poisson process with stochastic intensity). Write $K(t)$ for the number of mutations up to time $t \geq 0$, and write

$$T_i := \min\{t \geq 0 : K(t) = i\}$$

for the time of the ith mutation ($i \in \mathbb{N}$). Each mutation seeds a clone (subpopulation of mutants) which grows as a birth-death branching process with birth and death rates α_B and β_B. Write $Y_i(t)$ for the size of the ith clone time t after its initiation, and set $Y_i(0) = 1$. It is assumed that the $Y_i(\cdot)$ are independent of each other, and of $A(\cdot)$ and $K(\cdot)$.

The total mutant population size at time t is

$$B(t) = \sum_{i=1}^{K(t)} Y_i(t - T_i).$$

Note that the process counting the number of cells, $(A(t), B(t))_{t\geq 0}$, is a Markov process on $\mathbb{N}_0 \times \mathbb{N}_0$, with transition rates

$$(i,j) \mapsto \begin{cases} (i+1,j), & \text{rate } i\alpha_A; \\ (i-1,j), & \text{rate } i\beta_A; \\ (i,j+1), & \text{rate } i\nu + j\alpha_B; \\ (i,j-1), & \text{rate } j\beta_B. \end{cases}$$

We are interested in the process at a fixed time t, and at the random times

$$\sigma_n := \min\{t \geq 0 : A(t) + B(t) \geq n\}$$

and

$$\tau_n := \min\{t \geq 0 : A(t) \geq n\},$$

for $n \in \mathbb{N}$. Trivially, $\sigma_n \leq \tau_n$.

The deterministic time t is especially relevant for in vitro experimental settings, where the age of the process is known. But what about a tumour? Its age is unknown while its size can be measured. Thus one might consider the random times at which the population reaches a given size. An application of the model is the emergence of drug resistance in a tumour. Here type A and B cells represent drug sensitive and resistant cells respectively. In this case the times σ_n are relevant. Another interpretation of the model is metastasis. Here type A and B cells make up the primary and secondary tumours. In this case the times τ_n are relevant.

Write $\lambda_A = \alpha_A - \beta_A$ and $\lambda_B = \alpha_B - \beta_B$ for the growth rates of the wildtype and mutant cells. We shall only be concerned with the case of supercritical wildtype growth, $\lambda_A > 0$.

Remark 2.1.1. *Kendall's model neglects the event that a wildtype cell divides to produce two mutant cells ($A \to BB$), and neglects the event that a mutant cell divides to produce wildtype cells ($B \to BA$ and $B \to AA$). In Chapter 3 such events are allowed.*

2.2 Large-time and large-population limits

How does the mutant growth trajectory appear at large times? This question is mostly already well understood. Durrett and Moseley [13] study the case $\lambda_A < \lambda_B$. Janson [21] studies a broad class of urn models, which encompasses Kendall's model in the case $\lambda_A > \lambda_B$. We do not present results as detailed as Janson's. Our aim for this section is not to offer a comprehensive study, but rather bring together basic results which give valuable insight.

First let's note the long-term wildtype growth rate. According to Lemma 1.2.3,

$$\lim_{t\to\infty} e^{-\lambda_A t} A(t) = W \tag{2.1}$$

almost surely, where

$$W \stackrel{d}{=} \sum_{i=1}^{A(0)} \chi_i \psi_i,$$

and the $\chi_i \sim \text{Bernoulli}(\lambda_A/\alpha_A)$ and the $\psi_i \sim \text{Exponential}(\lambda_A/\alpha_A)$ are independent.

As for the long-term mutant growth rate, there is a trichotomy, depending on the relative fitness of wildtype and mutant cells. Part 1 of Theorem 2.2.1 is a special case of [21, Theorem 3.1], and part 3 is [13, Theorem 2].

Theorem 2.2.1 (Large time limit)**.** *The following limits hold almost surely.*

1. *For* $\lambda_A > \lambda_B$,
$$\lim_{t\to\infty} e^{-\lambda_A t} B(t) = \frac{\nu}{\lambda_A - \lambda_B} W.$$
2. *For* $\lambda_A = \lambda_B$,
$$\lim_{t\to\infty} t^{-1} e^{-\lambda_A t} B(t) = \nu W.$$
3. *For* $\lambda_A < \lambda_B$,
$$\lim_{t\to\infty} e^{-\lambda_B t} B(t) = V.$$

The limit random variable W comes from (2.1). The limit random variable V is $[0,\infty)$-valued with mean

$$\mathbb{E}[V] = \frac{A(0)\nu}{\lambda_B - \lambda_A}.$$

The full distribution of V is given in [3, Section 4.3], which we do not state here for the sake of brevity.

For $\lambda_A \geq \lambda_B$, conditioned on wildtype non-extinction, any individual clone ultimately makes up zero proportion of the mutant population. That is to say, conditioned on $W > 0$ (Lemma 1.2.4),

$$\lim_{t\to\infty} \frac{Y_i(t - T_i)}{B(t)} = 0$$

almost surely. We say that the mutant population is driven by the wildtype growth. This is seen in the limit random variables' dependence on W.

For $\lambda_A < \lambda_B$, early arriving clones make an important contribution to the mutant population. Conditioned on $W > 0$,

$$\lim_{t\to\infty} \frac{Y_i(t - T_i)}{B(t)} = \frac{X_i e^{-\lambda_B T_i}}{V}$$

almost surely. Note that if $W > 0$, then $V > 0$ [13]. The X_i are i.i.d. with distribution $\chi^B \psi^B$, where $\chi^B \sim \text{Bernoulli}(\lambda_B/\alpha_B)$ and $\psi^B \sim \text{Exponential}(\lambda_B/\alpha_B)$ are independent. We say that the mutant population is driven by the clone growth.

To see the asymptotic behaviour of the number of mutations, simply consider $\alpha_B = \beta_B = 0$ in Theorem 2.2.1:

$$\lim_{t\to\infty} e^{-\lambda_A t} K(t) = \frac{\nu}{\lambda_A} W$$

almost surely.

As corollaries to Theorem 2.2.1 we obtain large population limits. Note that conditioned on $W > 0$, $\lim_{n\to\infty} \tau_n = \lim_{n\to\infty} \sigma_n = \infty$ almost surely.

Corollary 2.2.2 (Large wildtype population limit)**.** *Conditioned on $W > 0$, the following limits hold almost surely.*

1. *For $\lambda_A > \lambda_B$,*
$$\lim_{n\to\infty} n^{-1} B(\tau_n) = \frac{\nu}{\lambda_A - \lambda_B}.$$
2. *For $\lambda_A = \lambda_B$,*
$$\lim_{n\to\infty} (n \log(n))^{-1} B(\tau_n) = \frac{\nu}{\lambda_A}.$$
3. *For $\lambda_A < \lambda_B$,*
$$\lim_{n\to\infty} n^{-\lambda_B/\lambda_A} B(\tau_n) = V W^{-\lambda_B/\lambda_A}.$$

Corollary 2.2.3 (Large total population limit)**.** *Conditioned on $W > 0$, the following limits hold almost surely.*

1. *For $\lambda_A > \lambda_B$,*
$$\lim_{n\to\infty} n^{-1} B(\sigma_n) = \frac{\nu}{\lambda_A - \lambda_B + \nu}.$$
2. *For $\lambda_A = \lambda_B$,*
$$\lim_{n\to\infty} n^{-1} \log(n)(n - B(\sigma_n)) = \frac{\lambda_A}{\nu}.$$
3. *For $\lambda_A < \lambda_B$,*
$$\lim_{n\to\infty} n^{-\lambda_A/\lambda_B}(n - B(\sigma_n)) = V^{-\lambda_A/\lambda_B} W.$$

Note that $n - B(\sigma_n) = A(\sigma_n)$. In case 1, the wildtype and mutant cells come to coexist in a constant ratio. In cases 2 and 3, the mutant cells eventually dominate the overall population, with

$$\lim_{n\to\infty} n^{-1} B(\sigma_n) = 1 \tag{2.2}$$

almost surely.

The long-term limiting trajectory of the population growth is perhaps a natural starting point in an investigation of the model's behaviour. But are these

long-term limits biologically relevant? At first glance it appears so. In many applications, the process runs for a long time until there are many cells. Having said that, it is often the case that, as well as there being many cells, the mutation rate is extremely small. In the next two sections, it is additionally considered that the mutation rate is small. Entirely different limiting behaviour is observed.

2.3 Large-population small-mutation limit

A tumour may comprise around 10^9 cells upon detection, with mutation rates per base pair per cell division estimated as 5×10^{-10} in colorectal cancer [23], for example. Hence a biologically relevant limit can be found by taking the final population size to infinity and the mutation rate to zero, while keeping their product finite.

For notation, let's include a superscipt ν in the random variables to denote their dependence on the mutation rate ν. That is, write $B^\nu(\cdot)$, $K^\nu(\cdot)$, T_i^ν, and σ_n^ν. Note that the random variables $A(\cdot)$, $Y_i(\cdot)$, and τ_n do not depend on ν.

First a connection is seen between the times τ_n and σ_n^ν.

Proposition 2.3.1. *Taking $\nu \to 0$ and $n\nu \to \theta < \infty$,*

$$(\tau_n - \sigma_n^\nu | \tau_n < \infty) \to 0$$

in distribution.

To clarify the meaning of the notation in Proposition 2.3.1, $(\tau_n - \sigma_n^\nu | \tau_n < \infty)$ is the random variable $\tau_n - \sigma_n^\nu$ conditioned on the event that $\tau_n < \infty$. Such conditioning will be seen throughout this section. In applications the process is observed when n cells are reached, so conditioning that n cells are reached (rather than extinction of the population first) is clearly appropriate. Proposition 2.3.1 says, in words, that the time difference between the wildtype and total population sizes reaching n is negligible. As a consequence, all results of this section will hold both in terms of the wildtype population and total population sizes. That is to say, using τ_n or σ_n^ν as the time variable will yield the same distributions in the limit. To save writing each result twice, we introduce the sequence (ρ_n^ν), which may refer to (τ_n) or (σ_n^ν).

The next result underlies all subsequent results of this section, saying that the times of mutation centered about (ρ_n^ν) converge to a Poisson process.

Theorem 2.3.2 (Mutations times)**.** *Taking $\nu \to 0$ and $n\nu \to \theta < \infty$,*

$$(K^\nu(\rho_n^\nu + t) | \rho_n^\nu < \infty) \to K^*(t)$$

in finite dimensional distributions. $K^*(t)$ is a Poisson process on $\mathbb{R}$ with intensity $\theta e^{\lambda_A t}$.

A direct consequence of Theorem 2.3.2 is that for each $i \in \mathbb{N}$, as $\nu \to 0$ and $n\nu \to \theta$,

$$(T_i^\nu - \rho_n^\nu | \rho_n^\nu < \infty) \to T_i^* := \min\{t \in \mathbb{R} : K^*(t) = i\}$$

in distribution. In particular, the time of first mutation T_1^* has Gumbel distribution:

$$\mathbb{P}[T_1^* \geq t] = \exp\left(-\frac{\theta}{\lambda_A} e^{\lambda_A t}\right).$$

Next we look at the clone sizes.

Theorem 2.3.3. *Taking $\nu \to 0$ and $n\nu \to \theta < \infty$,*

$$\left(\sum_{i=1}^{K^\nu(\rho_n^\nu)} \delta_{Y_i(\rho_n^\nu - T_i^\nu)} \Big| \rho_n^\nu < \infty\right) \to \sum_{i=1}^{K^*(0)} \delta_{Y_i^*}$$

in distribution (on the space of measures on $[0,\infty)$ equipped with the vague topology), where the Y_i^* are i.i.d. with $Y_1^* \stackrel{d}{=} Y_1(\xi)$ and $\xi \sim$Exponential(λ_A), and the Y_i^* are independent of $K^*(0)$.

Let's explain the limit of Theorem 2.3.3. According to Theorem 2.3.2, at time ρ_n^ν there will be approximately $K^*(0)$ clones, and the unordered ages of these clones follow the distribution of ξ. Thus the Y_i^* are the (unordered) clone sizes. From [4, page 109],

$$\mathbb{E}[z^{Y_i(t)}] = \frac{\beta_B(z-1) - e^{-\lambda_B t}(\alpha_B z - \beta_B)}{\alpha_B(z-1) - e^{-\lambda_B t}(\alpha_B z - \beta_B)}, \tag{2.3}$$

and so

$$\begin{aligned} r(z): &= \mathbb{E}[z^{Y_i^*}] \\ &= \int_0^\infty \mathbb{E}[z^{Y_i(t)}]\lambda_A e^{-\lambda_A t} dt \\ &= 1 - (1-q_B) F\left(\begin{matrix}1, \lambda_A/\lambda_B \\ 1+\lambda_A/\lambda_B\end{matrix}; \frac{q_B - z}{1-z}\right). \end{aligned} \tag{2.4}$$

The function F is Gauss's hypergeometric function , and $q_B = \beta_B/\alpha_B$, which is a clone's ultimate extinction probability if $q_B \leq 1$. The third equality of (2.4) can be seen by making a change of variable $s = e^{-\lambda_B t}$, and then using a standard integral representation for F (for example [24, C.8]).

Next we see the number of mutants.

Corollary 2.3.4 (Number of mutants)**.** *Taking $\nu \to 0$ and $n\nu \to \theta < \infty$,*

$$(B^\nu(\rho_n^\nu) | \rho_n^\nu < \infty) \to B^* := \sum_{i=1}^{K^*(0)} Y_i^*$$

in distribution.

Clearly B^* of Corollary 2.3.4 is a compound Poisson random variable, and

has generating function

$$\mathbb{E}[z^{B^*}] = \exp\left(\frac{\theta}{\lambda_A}(r(z)-1)\right). \tag{2.5}$$

This recovers recent results of Kessler and Levine [27] who provided a heuristic derivation of this expression, and Keller and Antal [24] who derived it for a deterministic exponentially growing wildtype population. Its large θ limit appeared in Durrett and Moseley [13] for $\lambda_A < \lambda_B$ (see [24] for a discussion).

Remark 2.3.5. *For $\lambda_B > 0$, the generating functions (2.4) and (2.5) yield power law tails:*

$$\lim_{k\to\infty} k^{1+\lambda_A/\lambda_B}\mathbb{P}[Y_i(\xi_i)=k] = \frac{\lambda_A}{\lambda_B}(1-q_B)^{1-\lambda_A/\lambda_B}\Gamma(1+\lambda_A/\lambda_B)$$

and

$$\lim_{k\to\infty} k^{1+\lambda_A/\lambda_B}\mathbb{P}[B^*=k] = \frac{\theta}{\lambda_B}(1-q_B)^{1-\lambda_A/\lambda_B}\Gamma(1+\lambda_A/\lambda_B),$$

which are given in [34, 24, 27].

The following definition will be used later in the thesis.

Definition 2.3.6. *A random variable with generating function (2.5) is said to have Luria-Delbrück distribution with parameters*

$$(\alpha_B/\lambda_A, \beta_B/\lambda_A, \theta/\lambda_A).$$

A special case of the Luria-Delbrück distribution, with parameters $(1,0,\theta)$, recovers the distribution derived by Lea and Coulson [31]: (2.5) reduces to

$$\mathbb{E}[z^{B^*}] = (1-z)^{\theta(z^{-1}-1)},$$

and the power-law tail is

$$\lim_{k\to\infty} k^2\mathbb{P}[B^*=k] = \theta.$$

Of potential interest is the number of clones of a given size, perhaps above some lower limit for reliable detection. Let I be a subset of $\mathbb{N}_0$. Consider

$$C_I^\nu(t) = \sum_{i=1}^{K^\nu(t)} 1_{\{Y_i(t-T_i^\nu)\in I\}}(t),$$

giving the number of clones whose size is in I at time t.

Corollary 2.3.7 (Number of clones of a given size)**.** *Taking $\nu \to 0$ and $n\nu \to \theta < \infty$,*

$$(C_I^\nu(\rho_n^\nu)|\rho_n^\nu < \infty) \to C_I^* \sim \text{Poisson}\left(\frac{\theta}{\lambda_A}\mathbb{P}[Y_1^* \in I]\right)$$

in distribution.

Next consider

$$M^\nu(t) = \max_{1 \le i \le K^\nu(t)} Y_i(t - T_i^\nu),$$

giving the size of the largest clone at time t.

Corollary 2.3.8 (Size of largest clone)**.** *Taking $\nu \to 0$ and $n\nu \to \theta < \infty$,*

$$(M^\nu(\rho_n^\nu) | \rho_n^\nu < \infty) \to M^*$$

in distribution, where $\mathbb{P}[M^* \le k] = \exp\left(-\frac{\theta}{\lambda_A}\mathbb{P}[Y_1^* > k]\right)$.

For an example let's consider the simplest choice of mutant cell growth: $\beta_B = 0$ and $\alpha_B = \lambda_A$. The number of clones above size k is

$$C^*_{\{i \in \mathbb{N}: i \ge k\}} \sim \text{Poisson}\left(\frac{\theta}{\lambda_A k}\right).$$

The size of the largest clone is

$$\mathbb{P}[M^* \le k] = \exp\left(-\frac{\theta}{\lambda_A(k+1)}\right).$$

Remark 2.3.9. *In this section we have considered a limit in which the product of the population size and mutation rate, $\theta = n\nu$, remains finite. It should be noted that alternative limits are also possible here. For example, Kessler and Levine [26] investigate large θ. In a different twist, Hamon and Ycart [19, Theorem 1.1] take the initial population size to infinity, the time of measurement to infinity, and the mutation rate to zero.*

2.4 Large-time small-mutation limit

Here we investigate results similar to Section 2.3, but with a view to approximating the process at a fixed time rather than population size. The time t is taken to infinity and the mutation rate ν to zero, with $\nu e^{\lambda_A t}$ converging. The superscript ν notation of Section 2.3 is used.

Theorem 2.4.1 (Mutation times)**.** *Taking $\nu \to 0$ and $t \to \infty$ with $\nu e^{\lambda_A t} \to \eta < \infty$,*

$$K^\nu(t_n + t) \to K^\circ(t)$$

in finite dimensional distributions. $K^\circ(t)$ is a Cox process on $\mathbb{R}$ with intensity $W\eta e^{\lambda_A t}$, where W is distributed as (2.1).

A direct consequence of Proposition 2.4.1 is that for each $i \in \mathbb{N}$,

$$T_i^\nu - t_n \to T_i^\circ := \min\{t \in \mathbb{R} : K^\circ(t) = i\} \tag{2.6}$$

in distribution. (If the reader is wondering what (2.6) means when the minimum is taken over an empty set, say that $\min \emptyset = \infty$, although this doesn't matter for subsequent results.)

Theorem 2.4.2. *Taking $\nu \to 0$ and $t \to \infty$ with $\nu e^{\lambda_A t} \to \eta < \infty$,*

$$\sum_{i=1}^{K^\nu(t)} \delta_{Y_i(t-T_i^\nu)} \to \sum_{i=1}^{K^\circ(0)} \delta_{Y_i^*}$$

in distribution, where the Y_i^* are given in Theorem 2.3.3 and are independent of $K^\circ(0)$.

Corollary 2.4.3 (Number of mutants)**.** *Taking $\nu \to 0$ and $t \to \infty$ with $\nu e^{\lambda_A t} \to \eta < \infty$,*

$$B^\nu(t) \to B^\circ = \sum_{i=1}^{K^\circ(0)} Y_i^*,$$

in distribution.

The generating function of B° is

$$\begin{aligned} \mathbb{E}[z^{B^\circ}] &= \mathbb{E}\left[\exp\left(\frac{W\eta}{\lambda_A}(r(z)-1)\right)\right] \\ &= \left(\frac{\lambda_A^2 - \beta_A\eta(r(z)-1)}{\lambda_A^2 - \alpha_A\eta(r(z)-1)}\right)^{A(0)}, \end{aligned} \tag{2.7}$$

where $r(z)$ is the clone size generating function, given by (2.4).

Remark 2.4.4. *For $\lambda_B > 0$, the generating function (2.7) yields the same power-law tail as (2.4) and (2.5) (see Remark 2.3.5):*

$$\lim_{k\to\infty} k^{1+\lambda_A/\lambda_B}\mathbb{P}[B^\circ = k] = \frac{A(0)\eta}{\lambda_B}(1-q_B)^{1-\lambda_A/\lambda_B}\Gamma(1+\lambda_A/\lambda_B).$$

The number of clones of a given size and the size of the largest clone can be determined in the large time small mutation limit in a similar manner to the large population small mutation limit.

Finally, we comment that the large time small mutation limit justifies a common approximation of Kendall's model, in which the wildtype population grows as $(We^{\lambda_A t})_{t\in\mathbb{R}}$. Here $B^\circ(\cdot)$ corresponds to $Z_1^*(\cdot)$ defined in [13], for example.

2.5 Exact results

Limit results simplify matters, offering clear insights into the model's behaviour. However limits make sense only in certain parameter regimes which may not always be appropriate for an application. Exact results, on the other hand, are complex and sometimes intractable but make sense for all parameter regimes. Some exact results are presented in this section.

What happens at a fixed wildtype population size? We are able to give the distribution of $B(\tau_n)$ in the special case of no wildtype cell death. (We also assume that the process begins with one wildtype cell, although it shouldn't be difficult to generalise this.)

Proposition 2.5.1. *For $A(0) = 1$ and $\beta_A = 0$,*

$$B(\tau_n) \stackrel{d}{=} \sum_{i=1}^{n-1} \sum_{j=1}^{K_i(\xi_i)} Y_{i,j}(U_{i,j}\xi_i), \tag{2.8}$$

where $(K_i(t))_{t\geq 0}$ are Poisson processes with intensity ν, $Y_{i,j}(\cdot) \stackrel{d}{=} Y_i(\cdot)$, $\xi_i \sim$Exponential(α_A), and $U_{i,j} \sim$Uniform[0,1], which are all independent.

To interpret (2.8), let's consider a randomly selected type A cell, labelled i, of the $n-1$ cells present just before time τ_n. The cell has been alive for time ξ_i, and initiated $K_i(\xi_i)$ mutant clones, with mutation times $(1-U_{i,j})\xi_i$ for $j = 1, 2, .., K_i(\xi_i)$. The clone sizes are $Y_{i,j}(U_{i,j}\xi_i)$.

The mean number of mutant cells at time τ_n is

$$\mathbb{E}[B(\tau_n)] = \begin{cases} \frac{(n-1)\nu}{\alpha_A - \lambda_B}, & \lambda_B < \alpha_A; \\ \infty, & \lambda_B \geq \alpha_A. \end{cases}$$

The generating function of $B(\tau_n)$ is

$$\begin{aligned} \mathbb{E}[z^{B(\tau_n)}] &= \left[\int_0^\infty \alpha_A e^{-\alpha_A t} \exp\left(\nu t \int_0^1 \mathbb{E}[z^{Y_{i,j}(ut)}] - 1 du\right) dt\right]^{n-1} \\ &= \left[\frac{1}{1 + \frac{\lambda_B \nu}{\alpha_A \alpha_B}} F\left(\begin{matrix} 1, \nu/\alpha_B \\ 1 + \nu/\alpha_B + \alpha_A/\lambda_B \end{matrix}; \frac{q_B - z}{q_B - 1}\right)\right]^{n-1}, \end{aligned}$$

where $\mathbb{E}[z^{Y_{i,j}(ut)}]$ is given by (2.3). The computation is lengthy but straightforward; one can apply the integral expression [24, C.8] for the hypergeometric function, and the identity [24, C.10]. As in Remarks 2.3.5 and 2.4.4, for $\lambda_B > 0$,

$$\lim_{k\to\infty} k^{1+\alpha_A/\lambda_B} \mathbb{P}[B(\tau_n) = k] \in (0, \infty) \tag{2.9}$$

exists. The limit can be obtained using the method of [24, Section 6] (which is based on [15]), but is too cumbersome to include here. Power-law tails have often appeared in two-type branching processes, but were generally considered to be an artefact of approximation [13, 41].

Remark 2.5.2. *Contrary to (2.9), moments of $B(\tau_n)$ are finite in the standard semi-deterministic version of the model (e.g. [31] and [24]).*

What about a fixed wildtype population size? Next, specialising further to neglect both wildtype and mutant death, we connect the distributions of the $B(\sigma_n)$ and $B(\tau_n)$.

Lemma 2.5.3. *For $\beta_A = \beta_B = 0$, and integers $0 \leq k < n$,*

$$\mathbb{P}[B(\sigma_n) \leq k] = \mathbb{P}[B(\tau_{n-k}) \leq k].$$

A similar result was given by Janson [22, Lemma 9.1] for a different class of urn models. Although Lemma 2.5.3 can be combined with Proposition 2.5.1

to determine the distribution of $B(\sigma_n)$, it does not seem likely that a tractable explicit expression can be obtained in general. However, for neutral mutations, Angerer was able to solve a recursion for the probabilities $\mathbb{P}[B(\sigma_n) = k]$ [2, Corollary 2.2].

Proposition 2.5.4 (Angerer). *For $A(0) = 1$, $\alpha_A + \nu = \alpha_B$ and $\beta_A = \beta_B = 0$,*

$$\mathbb{P}[B(\sigma_n) = k] = \sum_{i=1}^{n-k} (-1)^{n-i} \binom{n-k-1}{i-1} \binom{i\frac{\alpha_A}{\alpha_B} - 1}{n-1}.$$

Finally, let's remark upon a fixed time t.

Remark 2.5.5. *Antal and Kaprivsky [3] solved the Kolmogorov equations to determine the joint distribution of $(A(t), B(t))$. For brevity we do not restate their result here.*

2.6 Multiple sites and the site frequency spectrum

In this section we adapt Kendall's model and the results, specialising the setting in one sense while generalising in another. The specialisation is to neglect selection, which means that mutations have no effect on division and death rates. The generalisation is to consider mutations at multiple sites on the genome. This section can be thought of as a warm-up for Chapter 3.

Let's introduce the model. The overall population $(C(t))_{t\geq 0}$ grows as a birth-death branching process. Cells divide and die at rates α and β, where $\alpha > \beta$. Each cell is labelled by some sequence $(v_i)_{i\in\mathcal{S}} \in \{A, B\}^{\mathcal{S}}$, where $\mathcal{S}$ is a finite set denoting genetic sites. Here $v_i = A$ or $v_i = B$ means that site i is not mutated or mutated respectively. Suppose that a cell with sequence $(v_i)_{i\in\mathcal{S}}$ divides. It is replaced by two daughter cells with sequences $(V_i^1)_{i\in\mathcal{S}}$ and $(V_i^2)_{i\in\mathcal{S}}$, where for each $i \in \mathcal{S}$

$$(V_i^1, V_i^2) = \begin{cases} (v_i, v_i), & \text{probability } 1-\mu; \\ (B, v_i), & \text{probability } \mu/2; \\ (v_i, B), & \text{probability } \mu/2. \end{cases}$$

The (V_i^1, V_i^2) are assumed to be independent over $i \in \mathcal{S}$ and over different cell divisions (although the independence over i is not needed for this section). Assume that the initial cells have no mutations, that is, they all have sequences $(A)_{i\in\mathcal{S}}$.

Remark 2.6.1. *Just as in Kendall's model, slightly unnaturally, the event that a cell division sees two daughter cells mutate at the same genetic site is neglected. Backward mutations are also neglected. These assumptions simplify matters considerably, but will be removed in Chapter 3.*

At time $t \geq 0$, let's say that the cells' genetic sequences are $(v_i^{t,r})_{i\in\mathcal{S}}$ for $r = 1, .., C(t)$. This is the entirety of genetic information. However genetic data

is not usually so detailed. Mutation frequency data, like that of Figure 1.1, does not offer information on the level of single cells. It only keeps track of mutation frequencies at each site. Write

$$A_i(t) = |\{r \in \{1, .., C(t)\} : v_i^{t,r} = A\}|$$

and

$$B_i(t) = |\{r \in \{1, .., C(t)\} : v_i^{t,r} = B\}|$$

for the number of wildtype and mutant cells with respect to site $i \in \mathcal{S}$ at time t.

Remark 2.6.2. *Viewing the process at a single site recovers Kendall's model. Setting* $\alpha\mu = \nu$, $\alpha(1-\mu) = \alpha_A$, *and* $\beta = \beta_A$,

$$(A_i(t), B_i(t))_{t\geq 0} \stackrel{d}{=} (A(t), B(t))_{t\geq 0}$$

for each $i \in \mathcal{S}$.

If one is not concerned with the identity of sites, then a further reduction of information is suggested. The *site frequency spectrum* is a standard summary statistic of genetic data. It is defined as the number of sites who see mutations in a given number of cells, that is

$$|\{i \in \mathcal{S} : B_i(t) = k\}|$$

for $k \in \mathbb{N}_0$. By Remark 2.6.2 and linearity of expectation, the mean site frequency spectrum is determined by

$$\mathbb{E}|\{i \in \mathcal{S} : B_i(t) = k\}| = |\mathcal{S}|\mathbb{P}[B(t) = k]. \tag{2.10}$$

Given the site frequency spectrum's importance in data, (2.10) deserves to be emphasised. It is wonderfully simple, almost trivial: the mean site frequency spectrum is characterised by a single site's mutation frequency distribution. Now let's apply the results of previous sections.

First consider the time $\sigma_n = \min\{t \geq 0 : C(t) \geq n\}$ when the population size reaches n. For $\beta = 0$ and $C(0) = 1$, the mean site frequency spectrum is

$$\mathbb{E}|\{i \in \mathcal{S} : B_i(\sigma_n) = k\}| = |\mathcal{S}| \sum_{i=1}^{n-k} (-1)^{n-i} \binom{n-k-1}{i-1} \binom{i(1-\mu)-1}{n-1},$$

by Proposition 2.5.4. As for seeing the mean site frequency spectrum at a fixed time t, one can use Antal and Krapivsky's result [3] (Remark 2.5.5). What about the long term behaviour of the site frequency spectrum? The number of sites which are mutated in a given number of cells converges to zero: for any k,

$$\lim_{t\to\infty} |\{i \in \mathcal{S} : B_i(t) = k\}| = 0$$

almost surely. And all sites are eventually mutated in at least proportion $x \in$

$(0,1)$ of the population:

$$\lim_{n\to\infty} |\{i \in \mathcal{S} : B_i(\sigma_n) \geq xn\}| = |\mathcal{S}|$$

almost surely, due to (2.2).

Next we look at the small mutation limits. Include a superscript μ in the notation to denote the dependence on the mutation rate: write $A_i^\mu(\cdot)$, $B_i^\mu(\cdot)$, $C^\mu(\cdot)$, and σ_n^μ. Taking $\mu \to 0$ and $n \to \infty$ with $\mu\alpha n \to \theta < \infty$,

$$\mathbb{E}\left[|\{i \in \mathcal{S} : B_i^\mu(\sigma_n^\mu) = k\}| \middle| \sigma_n^\mu < \infty\right] \to |\mathcal{S}|\mathbb{P}[B^* = k], \tag{2.11}$$

where B^* is distributed according to (2.5) with $\alpha_A = \alpha_B = \alpha$ and $\beta_A = \beta_B = \beta$. Similarly, taking $\mu \to 0$ and $t \to \infty$ with $\mu\alpha e^{\lambda_A t} \to \eta < \infty$,

$$\mathbb{E}|\{i \in \mathcal{S} : B_i^\mu(t) = k\}| \to |\mathcal{S}|\mathbb{P}[B^\circ = k], \tag{2.12}$$

where B° is distributed according to (2.7) with $\alpha_A = \alpha_B = \alpha$ and $\beta_A = \beta_B = \beta$.

Remark 2.6.3. *The limits (2.11) and (2.12) mirror Corollaries 2.3.4 and 2.4.3 respectively. However there is a technical discrepancy in the reflection: unlike $A(\cdot)$ in the two-type model, the processes $A_i^\mu(\cdot)$ depend on the mutation rate. Therefore (2.11) and (2.12) cannot be deduced directly from 2.3.4 and 2.4.3. Rather than detail the proofs of (2.11) and (2.12), we wait to Chapter 3 for analogous results in an extended model.*

Remark 2.6.4. *Our approximations (2.11) and (2.12) for the mean site frequency spectrum have power-law tails:*

$$\lim_{k\to\infty} k^2|\mathcal{S}|\mathbb{P}[B^* = k] = \frac{|\mathcal{S}|\theta}{\lambda}$$

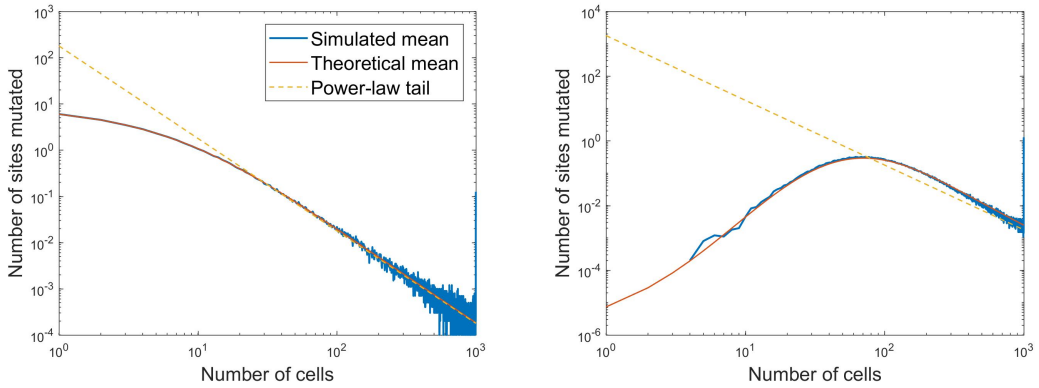
and

$$\lim_{k\to\infty} k^2|\mathcal{S}|\mathbb{P}[B^\circ = k] = \frac{|\mathcal{S}|\eta C(0)}{\lambda},$$

which are special cases of Remarks 2.3.5 and 2.4.4.

Since the size, rather than age, of a tumour can be observed, we are most interested in the large-population small-mutation limit. To give the reader an idea of its appearance, in Figure 2.1 the mean site frequency spectrum as given by (2.11) is plotted. The theoretical result is compared to simulations, with birth, death and scaled mutation rates taken from biological literature. In particular, we consider $\alpha = 0.25$ and $\beta = 0.18$ (per day), which were estimated in colorectal cancers by [8]. According to [23], θ may be of the order of α; we consider two different values for θ in this region. We take a relatively small population size of $n = 10^3$ and number of sites $|\mathcal{S}| = 50$, so that computation time is reasonable. It is expected that taking larger n and fixed θ will give an even closer fit between theory and simulations.

Figure 2.1: Simulated and theoretical expected site frequency spectrum, with $\alpha = 0.25$, $\beta = 0.18$, $|\mathcal{S}| = 50$, $C(0) = 1$, $n = 10^3$. Two different mutation rates are plotted: $\mu = 10^{-3}$ (left) and $\mu = 10^{-2}$ (right). The average has been taken over 10^4 simulations in each case.



2.7 Proofs

Proofs for Section 2.2

Proof of Theorem 2.2.1. For part 2, one needs to observe that

$$\left(e^{-\lambda_A t}B(t) - te^{-\lambda_A t}\nu A(t)\right)_{t\geq 0}$$

is a martingale with respect to the obvious filtration, and is bounded in L_2.

For part 1, the reader may refer to [21] for a full proof in a more general and notation-heavy setting. For the reader's convenience, we offer the essence of Janson's proof here. Crucially,

$$(M(t))_{t\geq 0} = \left(e^{-\lambda_B t}B(t) - \frac{\nu}{\lambda_A - \lambda_B}e^{-\lambda_B t}A(t)\right)_{t\geq 0}$$

is a martingale. Janson obtains bounds for the probabilities

$$\mathbb{P}\left[\sup_{t\in[n-1,n]}\left|e^{(\lambda_B-\lambda_A)t}M(t)\right| > \epsilon\right],$$

via Doob's martingale inequality, and then applies the Borel-Cantelli lemma. □

Proof of Corollary 2.2.3, part 2. First, rewrite

$$\begin{aligned}\frac{\log(n)A(\sigma_n)}{n} &= \frac{\log\left(A(\sigma_n)+B(\sigma_n)\right)A(\sigma_n)}{A(\sigma_n)+B(\sigma_n)}\\ &= \frac{1}{\sigma_n}\left[\log\left(\frac{A(\sigma_n)+B(\sigma_n)}{\sigma_n e^{\lambda_A\sigma_n}}\right)+\log(\sigma_n)+\lambda_A\sigma_n\right]\\ &\quad\times\frac{e^{-\lambda_A\sigma_n}A(\sigma_n)}{\sigma_n^{-1}e^{-\lambda_A\sigma_n}\left(A(\sigma_n)+B(\sigma_n)\right)}.\end{aligned}$$

Then apply Theorem 2.2.1 and (2.1), to take $n \to \infty$. □

The remaining parts of Corollaries 2.2.2 and 2.2.3 can be proven in a similar manner.

Proofs for Sections 2.3 and 2.4

We will construct the random variables in a way that allows weak convergence to be shown via almost sure convergence.

On a fresh probability space $(\Omega, \mathcal{F}, \mathbb{P})$ put the independent processes $(A(t))_{t\geq 0}$ and $(Y_i(t))_{t\geq 0}$ for $i \in \mathbb{N}$. Also put an independent Poisson counting process $(N(t))_{t\geq 0}$. Let $(\nu_n)_{n\in\mathbb{N}}$ be a sequence of mutation rates which satisfy

$$\lim_{n\to\infty} n\nu_n = \theta.$$

Now, for each $n \in \mathbb{N}$, define the mutation counting process

$$K^{\nu_n}(t) = N\left(\nu_n \int_0^t A(s)ds\right), \tag{2.13}$$

and the mutation times

$$T_i^{\nu_n} = \min\{t \geq 0 : K^{\nu_n}(t) = i\}$$

for $i \in \mathbb{N}$. Define the total number of mutants as

$$B^{\nu_n}(t) = \sum_{i=1}^{K^{\nu_n}(t)} Y_i(t - T_i^{\nu_n}).$$

Define the time at which n total cells are reached as

$$\sigma_n^{\nu_n} = \min\{t \geq 0 : A(t) + B^{\nu_n}(t) \geq n\}$$

and the time at which n wildtype cells are reached as

$$\tau_n = \min\{t \geq 0 : A(t) \geq n\}.$$

It is easy to see that the definitions here are equivalent to the definitions of Section 2.1. The difference is that extra information is included here, which gives the joint distribution of the random variables ranging over the sequence of mutation rates. The centrepiece of this joint distribution is (2.13), and it will soon become apparent that this choice allows relatively smooth proofs.

First let's prove one case of Proposition 2.3.2, but conditioning on the simpler event that $W > 0$ (where W is from (2.1)).

Lemma 2.7.1. *Condition on $\{W > 0\}$. As $n \to \infty$,*

$$K^{\nu_n}(\tau_n + t) = N\left(\int_{-\tau_n}^t \nu_n A(\tau_n + s)ds\right) \to N\left(\int_{-\infty}^t \theta e^{\lambda_A s}ds\right) =: K^*(t)$$

and

$$T_i^{\nu_n} - \tau_n \to T_i^* = \min\{t \in \mathbb{R} : K^*(t) = i\}$$

almost surely, for each $t \in \mathbb{R}$.

Proof. That $A(\cdot)$ is cadlag and satisfies (2.1), are enough to see that

$$\sup_{t \geq 0} \frac{A(t)}{e^{\lambda_A t}} < \infty,$$

and

$$\sup_{n \in \mathbb{N}} \frac{e^{\lambda_A \tau_n}}{A(\tau_n)} < \infty$$

almost surely. Now write

$$\nu_n A(\tau_n + t) = n\nu_n \frac{A(\tau_n + t)}{e^{\lambda_A(\tau_n + t)}} \frac{e^{\lambda_A \tau_n}}{A(\tau_n)} e^{\lambda_A t}.$$

It becomes apparent that

$$\lim_{n \to \infty} \nu_n A(\tau_n + t) = \theta e^{\lambda_A t},$$

and for all $t \in \mathbb{R}$

$$\sup_{n \in \mathbb{N}} \nu_n A(\tau_n + t) \leq L e^{\lambda_A t}$$

almost surely, for some positive random variable L. Then, using dominated convergence and the fact that $N(\cdot)$ is almost surely continuous at $\int_{-\infty}^{t} \theta e^{\lambda_A s} ds$, we are done. □

Lemma 2.7.2. *Condition on* $\{W > 0\}$.

$$\sup_{n \in \mathbb{N}} B^{\nu_n}(\sigma_n^{\nu_n}) < \infty$$

almost surely.

Proof. By Lemma 2.7.1,

$$\hat{K} := \sup_{n \in \mathbb{N}} K^{\nu_n}(\sigma_n^{\nu_n}) \leq \sup_{n \in \mathbb{N}} K^{\nu_n}(\tau_n) < \infty$$

and

$$\hat{T}_i := \sup_{n \in \mathbb{N}} (\tau_n - T_i^{\nu_n}) < \infty$$

almost surely. Then

$$\sigma_n^{\mu_n} - T_i^{\nu_n} \leq \tau_n - T_i^{\nu_n} \leq \hat{T}_i.$$

So

$$Y_i(\sigma_n^{\mu_n} - T_i^{\nu_n}) \leq \sup_{t \leq \hat{T}_i} Y_i(t) =: \hat{Y}_i < \infty,$$

(let's say that $Y_i(t) = 0$ for $t < 0$). Therefore

$$B^{\nu_n}(\sigma_n^{\nu_n}) = \sum_{i=1}^{K^{\nu_n}(\sigma_n^{\nu_n})} Y_i(\sigma_n^{\mu_n} - T_i^{\nu_n}) \leq \sum_{i=1}^{\hat{K}} \hat{Y}_i < \infty.$$

□

Lemma 2.7.3. *Condition on* $\{W > 0\}$. *As* $n \to \infty$,

$$\tau_n - \sigma_n^{\nu_n} \to 0$$

almost surely.

Proof. Consider a sequence of positive integers (a_n), such that

1. $\lim_{n\to\infty} a_n = \infty$, and
2. $\lim_{n\to\infty}(n - a_n)/n = 1$.

For example $a_n = \lfloor n^{1/2} \rfloor$ will do. Since

$$e^{\lambda_A(\tau_n - \tau_{n-a_n})} = \frac{We^{\lambda_A \tau_n}}{A(\tau_n)} \frac{A(\tau_n)}{n} \frac{n}{n-a_n} \frac{n-a_n}{A(\tau_{n-a_n})} \frac{A(\tau_{n-a_n})}{We^{\lambda_A \tau_{n-a_n}}}$$

converges to 1, we have that

$$\tau_n - \tau_{n-a_n}$$

converges to zero. By Lemma 2.7.2,

$$B^{\nu_n}(\sigma_n^{\nu_n}) \leq a_n$$

for sufficiently large n. For such n

$$A(\sigma_n^{\nu_n}) \geq n - a_n,$$

so

$$\sigma_n^{\nu_n} \geq \tau_{n-a_n},$$

and hence

$$0 \leq \tau_n - \sigma_n^{\nu_n} \leq \tau_n - \tau_{n-a_n}.$$

□

Lemma 2.7.4. *Condition on* $\{W > 0\}$. *As* $n \to \infty$,

$$K^{\nu_n}(\sigma_n^{\nu_n} + t) = K^*(t)$$

and

$$T_i^{\nu_n} - \sigma_n^{\nu_n} \to T_i^*$$

almost surely, for each $t \in \mathbb{R}$.

Proof. Combine Lemmas 2.7.1 and 2.7.3. □

Now that we have seen convergence of the mutation times, the convergence of clone sizes and the number of mutants follows immediately by continuity.

Lemma 2.7.5. *Condition on* $\{W > 0\}$. *For each* $i \in \mathbb{N}$,

$$\lim_{n\to\infty} Y_i(\tau_n - T_i^{\nu_n}) = \lim_{n\to\infty} Y_i(\sigma_n^{\nu_n} - T_i^{\nu_n}) = Y_i(-T_i^*)$$

almost surely.

Thanks to Lemma 2.7.5: the clone sizes converge when the event $\{W > 0\}$ is conditioned on. On the other hand, Theorem 2.3.3 and Corollaries 2.3.4, 2.3.7, and 2.3.8 say that the event $\{\rho_n^{\nu_n} < \infty\}$ is conditioned on. The same difference is seen between Lemmas 2.7.1 and 2.7.4 and Theorem 2.3.2, with the lemmas conditioning on $\{W > 0\}$ and the theorem conditioning on $\{\rho_n^{\nu_n} < \infty\}$. This means that the only remaining task for large-population small-mutation limit proofs is to show that it makes no difference to condition on $\{W > 0\}$ or $\{\rho_n^{\nu_n} < \infty\}$. The following two lemmas show exactly this.

Lemma 2.7.6. *Suppose that* $(E_n)_{n\in\mathbb{N}}$ *and* $(F_n)_{n\in\mathbb{N}}$ *are sequences of events, such that*

1. $\forall n \in \mathbb{N}(F_n \supset F_{n+1})$,
2. $\cap_{n\in\mathbb{N}} F_n = F$, *and*
3. $\mathbb{P}[F] > 0$.

Then

$$\lim_{n\to\infty} \mathbb{P}[E_n|F_n] = \lim_{n\to\infty} \mathbb{P}[E_n|F],$$

if it exists.

Proof. Write

$$\mathbb{P}[E_n|F_n] = \frac{\mathbb{P}[F]}{\mathbb{P}[F_n]}\mathbb{P}[E_n|F] + \frac{\mathbb{P}[E_n \cap F_n \backslash F]}{\mathbb{P}[F_n]},$$

and take $n \to \infty$. □

Lemma 2.7.7.

$$\{W > 0\} = \cap_{n\in\mathbb{N}}\{\tau_n < \infty\} = \cap_{n\in\mathbb{N}}\{\sigma_n^{\nu_n} < \infty\},$$

where W *is given by (2.1).*

Proof. That $\{W > 0\} \subset \cap_{n\in\mathbb{N}}\{\tau_n < \infty\} \subset \cap_{n\in\mathbb{N}}\{\sigma_n^{\nu_n} < \infty\}$ should be clear. We show that

$$\cap_{n\in\mathbb{N}}\{\sigma_n^{\nu_n} < \infty\} \subset \{W > 0\}.$$

Fix $\omega \in \{W = 0\} = \{\exists t \geq 0, A(t) = 0\}$. Then

$$\int_0^\infty A(t)dt < \infty.$$

So one can choose sufficiently large $n \in \mathbb{N}$ such that both

$$\nu_n \int_0^\infty A(t)dt < \sup\{t \geq 0 : N(t) = 0\},$$

and

$$\tau_n = \infty,$$

which means that

$$\omega \in \{\forall t \geq 0, K^{\nu_n}(t) = 0\} \cap \{\tau_n = \infty\} \subset \{\sigma_n^{\nu_n} = \infty\}.$$

□

Next let's look at Theorems 2.4.1 and 2.4.2, which are the large-time small-mutation limit results. The large-time small-mutation limit is essentially a much easier version of the large-population small-mutation limit. There is no need to condition on any event, and the times are deterministic rather than random.

Proof of Theorem 2.4.1. Let (t_n) be a sequence of times with

$$\lim_{n\to\infty} \nu_n e^{\lambda_A t_n} = \eta.$$

It needs to be shown that for each $t \in \mathbb{R}$

$$N\left(\int_{-t_n}^t \nu_n A(t_n + s)ds\right) \to N\left(\int_{-\infty}^t W\eta e^{\lambda_A s}ds\right)$$

almost surely, as $n \to \infty$. Indeed, writing

$$\nu_n A(t_n + s) = \nu_n e^{\lambda_A t_n} \frac{A(t_n + s)}{e^{\lambda_A(t_n+s)}} e^{\lambda_A s},$$

one sees that $\nu_n A(t_n + s)$ converges to the appropriate limit and is dominated by a multiple of $e^{\lambda_A s}$. □

Theorem 2.4.2 follows by continuity.

Proofs for Section 2.5

First let's state a classic result which can be found in [36].

Lemma 2.7.8. *Assume that $\beta_A = 0$. For each n, $(\tau_n - \tau_k)_{k=1}^{n-1}$ has the same distribution as a collection of $n-1$ i.i.d. Exponential(α_A) random variables, which are ordered by size.*

Proof of Proposition 2.5.1. For each $i \in \mathbb{N}$ let $(T_{i,j})_{j\in\mathbb{N}}$ be the occurrence times of a homogeneous Poisson process on $[0, \infty)$ with intensity ν. These are the mutation times corresponding to one particular wildtype cell present from time 0. Noting that

$$A(t) = \sum_{i=1}^{\infty} 1_{[\tau_i,\infty)}(t),$$

it is apparent that the mutation times of all wildtype cells are distributed according to

$$(\tau_i + T_{i,j})_{i,j\in\mathbb{N}}.$$

The number of mutants at time t is

$$B(t) \stackrel{d}{=} \sum_{i,j\in\mathbb{N}} 1_{\{t-\tau_i-T_{i,j}\geq 0\}} Y_{i,j}(t-\tau_i-T_{i,j}) = \sum_{i\in\mathbb{N}} 1_{\{t-\tau_i\geq 0\}} D_i(t-\tau_i),$$

where

$$D_i(t) = \sum_{j\in\mathbb{N}} 1_{\{t-T_{i,j}\geq 0\}} Y_{i,j}(t-T_{i,j}) \stackrel{d}{=} \sum_{j=1}^{K_i(t)} Y_{i,j}(U_{i,j}t).$$

The $D_i(\cdot)$ are i.i.d. Now, using Lemma 2.7.8,

$$B(\tau_n) \stackrel{d}{=} \sum_{i=1}^{n-1} D_i(\tau_n-\tau_i) \stackrel{d}{=} \sum_{i=1}^{n-1} D_i(\xi_i),$$

and by substituting $D_i(\cdot)$ the result is obtained.

□

Proof of Lemma 2.5.3. We will show that the events $\{B(\sigma_n)\leq k\}$ and $\{B(\tau_{n-k})\leq k\}$ are equal, using the monotonicity of $A(\cdot)$ and $B(\cdot)$ and the fact that $A(\sigma_n)+B(\sigma_n)=n$. First assume that $B(\sigma_n)\leq k$. Then $A(\sigma_n)\geq n-k$, so $\sigma_n\geq\tau_{n-k}$, and therefore $B(\tau_{n-k})\leq k$. Now assume that $B(\sigma_n)>k$. Then $A(\sigma_n)<n-k$, so $\sigma_n<\tau_{n-k}$, and hence $B(\tau_{n-k})>k$.

□

2.8 Discussion

Let's discuss a single application, which is the main application in mind for the thesis: genetic diversity in cancer. With the advent of next-generation DNA sequencing, vast quantities of cancer genomes have been sequenced. Data has been made publicly available through the Cancer Genome Atlas and International Cancer Genome Consortium, for example. Considerable efforts have been made in recent years to explain observed mutation patterns with mathematical models, and from the observed mutation patterns to infer the evolutionary history of tumours.

Striking examples are Williams et al. [39] and Bozic et al. [7], who consider deterministic and branching process models respectively. They both derive that the expected frequency of mutations occurring in x proportion of cells has density proportional to x^{-2} (away from 0). In [39], 323 out of 904 cancers considered are deemed to fit the x^{-2} power-law. In [7], 14 out of 42 cancers are deemed to fit the power-law.

The models of [39, 7, 35, 10] all used the infinite-sites assumption, which states that each site can mutate at most once over the lifetime of a tumour. Statistical analysis of cancer genomic data refutes this assumption [30]. Furthermore, we make a theoretical argument against the infinite-sites assumption in the branching process setting. According to Theorem 2.3.2, the number of times a particular site has mutated before the population size reaches n is approximately Poisson$(n\nu/\lambda_A)$. Therefore the infinite-sites simplification may be appropriate when $n\nu/\lambda_A$ is much smaller than 1. However [39] estimated effective mutation rates, ν/λ_A, of single base pairs to be in the region of $10^{-7} - 10^{-6}$. If a detected tumour comprises $10^8 - 10^9$ cells (e.g. [7]), then $n\nu/\lambda_A$ is not sufficiently small.

We have shown that the mean site frequency spectrum can be approximated by a well known generalisation of the Luria-Delbrück distribution. The distribution's x^{-2} tail agrees with theoretical predictions and data in [39, 7]. But, as seen in Figure 2.1, our predictions disagree at the lower end of the frequency spectrum. Due to unreliable data, [39, 7] did not make a model-data comparison for mutations occurring in less than 10% of cells.

Chapter 3 will see the model extended, the results deepened, and this discussion developed further.

Chapter 3

Multiple genetic sites

In Chapter 2 we explored a branching process model of a growing cell population with binary genetic information (which keeps track of each cell's mutational status at a particular genetic site). Then we briefly extended the model and some results to a sequence representation of genetic information (which keeps track of multiple genetic sites). Now in this chapter we redefine the sequence model in a more natural way, saying that each cell contains a sequence of the nucleotides A, C, G, and T, with all possible genetic transitions allowed at cell divisions. Furthermore we now considerably deepen the results of Chapter 2. With the increased depth however, our focus is also more narrow, considering only the limit as the final number of cells converges to infinity and the mutation rates converge to zero. Let's give an overview of this chapter's results.

- At a single site (nucleotide), the number of cells which are mutated converges to a Luria-Delbrück random variable, recovering Corollary 2.3.4. This means that, in the limit, the number of mutant cells is finite as opposed the infinite total number of cells. As for the rare event that the fraction of mutant cells exceeds a positive number, it is shown that the probability of this event scales with the mutation rate, recovering the power-law tail of the Luria-Delbrück distribution.

- Across multiple sites, the joint distribution of mutation frequencies reflects the evolutionary tree's random structure. Notably, independence transitions to dependence from small to large frequencies.

- Taking the number of sites to infinity (along with the final number of cells to infinity and the mutation rates to zero), the site frequency spectrum converges to a determinisitic limit at small frequencies and to a Cox process at large frequencies. The Cox process's random intensity measure is a function of the evolutionary tree's structure.

These results and their proofs are valuable on several fronts. First, the distribution of the site frequency spectrum allows enlightened data comparison (previous works only offer the expected site frequency spectrum [39, 7, 10]). Second, the infinite-sites assumption is exposed; widespread violations of the infinite-sites assumption are seen, but the impact of violations on mutation frequencies is limited. Third, the impact of site-specific mutation rates is made clear. Fourth,

the impact of selection is understood in a special case, and further research is suggested.

The chapter is structured as follows. In Section 3.1, the model in its most basic form is introduced (no cell death, no selection, homogeneous mutation rates). In Section 3.2, some notation and the parameter regime are introduced. In Section 3.3, small-frequency mutations are discussed. In Section 3.4, large-frequency mutations are discussed. In Section 3.5, the model and results are generalised to cell death, selection, and site-specific mutation rates, and some conjectures are posed. In Section 3.6, proofs are given. In Section 3.7, support for conjectures is given. In Section 3.8, mutation rates in a lung adenocarcinoma are estimated.

3.1 Model

Here the model is stated in its simplest form. It comprises two parts.

1. Population dynamics: Starting with one cell, cells divide according to the Yule process. That is to say, when there are $k \in \mathbb{N}$ cells, a cell is chosen uniformly at random to divide. The Yule process can equivalently be regarded as a continuous-time Markov process, with cells dividing independently at constant rate.

2. Genetic information: The set of nucleotides is

$$\mathcal{N} = \{A, C, G, T\}.$$

Each cell has a genome, which is a finite sequence of nucleotides. Genomes are elements of the set

$$\mathcal{G} = \mathcal{N}^{\mathcal{S}},$$

where $\mathcal{S}$ is a finite set denoting genetic sites. Suppose that a cell with genome $(v_i)_{i\in\mathcal{S}} \in \mathcal{G}$ divides. The cell is replaced by two daughter cells with genomes $(V_i^1)_{i\in\mathcal{S}}$ and $(V_i^2)_{i\in\mathcal{S}}$, where:

- The V_i^r are independent over $i \in \mathcal{S}$ and $r \in \{1, 2\}$. (It is also assumed that mutations are independent for different cell divisions.)
- $V_i^r = v_i$, with probability $1 - \mu/2$; or
  V_i^r is uniformly distributed on $\mathcal{N}\backslash\{v_i\}$, with probability $\mu/2$.

Remark 3.1.1. *The factor of* $1/2$ *in the mutation probability balances the fact that two new cells are produced at cell division. So the expected number of mutations per site per cell division is* μ.

Remark 3.1.2. *There is nothing special, mathematically, about the set of nucleotides* $\mathcal{N}$. *It can be replaced with any finite set.*

The model is generalised to cell death, selection, and heterogeneous mutation rates in Section 3.5.

3.2 Preliminaries

3.2.1 Notation

Consider a homogeneous mutation rate μ. Write $X_v^{n,\mu}$ for the number of cells with genome $v \in \mathcal{G}$ when there are n cells in total. Say that $u \in \mathcal{G}$ is the initial cell's genome, so $X_v^{1,\mu} = \delta_{u,v}$. A genetic site is said to be mutated if its nucleotide differs from that of the initial cell. Write

$$\mathcal{G}_{(i)} = \{v \in \mathcal{G} : v_i \neq u_i\}$$

for the set of genomes which are mutated at site $i \in \mathcal{S}$. Write

$$B_i^{n,\mu} = \sum_{v \in \mathcal{G}_{(i)}} X_v^{n,\mu} \tag{3.1}$$

for the number of cells which are mutated at site $i \in \mathcal{S}$ when there are n cells in total. The quantity (3.1) is the primary subject of our study.

3.2.2 Parameter regime

As discussed in Section 2.3, the number of cells in a detected tumour may be in the region of $n = 10^9$, whereas the point mutation rate per site per cell division is in the region of $\mu = 10^{-9}$. The number of base pairs in the human genome is around $|\mathcal{S}| = 3 \times 10^9$. Very roughly,

$$n \approx \mu^{-1} \approx |\mathcal{S}|.$$

Therefore we study the limits:

- $\mu \to 0$, $n\mu \to \theta < \infty$;
- $\mu \to 0$, $n\mu \to \theta < \infty$, $|\mathcal{S}| \to \infty$ (sometimes with $|\mathcal{S}|\mu \to \eta < \infty$).

These parameter regimes are not only relevant for cancer. The $n\mu \to \theta$ limit has long been popular in similar models of bacteria, beginning with [32, 31].

Remark 3.2.1. *In Chapter 2 the large-population small-mutation limit was written as $n\mu\alpha \to \theta$, as opposed to $n\mu \to \theta$ here. The difference is only for notational convenience.*

3.3 Small-frequency mutations

The majority of sites are mutated in only a small proportion of cells. Let's briefly explain. Late in the population growth trajectory there are many cells and many divisions. Hence there are many late-arising mutations, which do not have enough time to reach large frequencies. Moreover these mutations are likely arise on distinct branches of the evolutionary tree, which evolve independently.

The first result shows that the number of cells mutated at any particular site converges to a Luria-Delbrück random variable (see definition 2.3.6), and that sites see asymptotic independence. The single-site convergence recovers Theorem 2.3.4. The independence between sites is new.

Theorem 3.3.1. *As* $\mu \to 0$ *and* $n\mu \to \theta \in (0, \infty)$,

$$(B_i^{n,\mu})_{i \in \mathcal{S}} \to (B_i)_{i \in \mathcal{S}}$$

in distribution, where the B_i *are independent Luria-Delbrück random variables with parameters* $(1, 0, \theta)$.

The *site frequency spectrum* is defined by the number of sites which are mutated in k cells,

$$\left|\{i \in \mathcal{S} : B_i^{n,\mu} = k\}\right|,$$

for $k = 0, 1, .., n$. By Theorem 3.3.1 and linearity of expectation, the expected site frequency spectrum is

$$\mathbb{E}\left|\{i \in \mathcal{S} : B_i^{n,\mu} = k\}\right| \approx |\mathcal{S}| \mathbb{P}[B_i = k],$$

whose shape is simply the Luria-Delbrück distribution. More can be said. Thanks to the independence in Theorem 3.3.1, a law of large numbers for the site frequency spectrum is obtained.

Corollary 3.3.2. *As* $\mu \to 0$, $n\mu \to \theta \in (0, \infty)$, *and* $|\mathcal{S}| \to \infty$,

$$\frac{\left|\{i \in \mathcal{S} : B_i^{n,\mu} = k\}\right|}{|\mathcal{S}|} \xrightarrow{p} \mathbb{P}[B_i = k].$$

That is to say, if the genome is large then the site frequency spectrum's shape is almost deterministic, given by the Luria-Delbrück distribution.

Theorem 3.3.1 and Corollary 3.3.2 describe the 'bulk' of the mutation frequency distribution. They say that the bulk is located at a small frequency relative to the population size. For any $\epsilon > 0$, the proportion of sites which are mutated in fewer than proportion ϵ of cells is

$$\frac{\left|\{i \in \mathcal{S} : n^{-1} B_i^{n,\mu} < \epsilon\}\right|}{|\mathcal{S}|}, \tag{3.2}$$

which converges to 1 in the limits of Theorem 3.3.1 and Corollary 3.3.2. Next we explain what happens away from the bulk, for those unusual sites which are mutated in a large proportion of cells.

3.4 Large-frequency mutations

A minority of sites are mutated in a large proportion of cells. These sites have a non-trivial dependency structure which is intimately related to the evolutionary tree. Let's briefly explain. Early in the population growth trajectory there

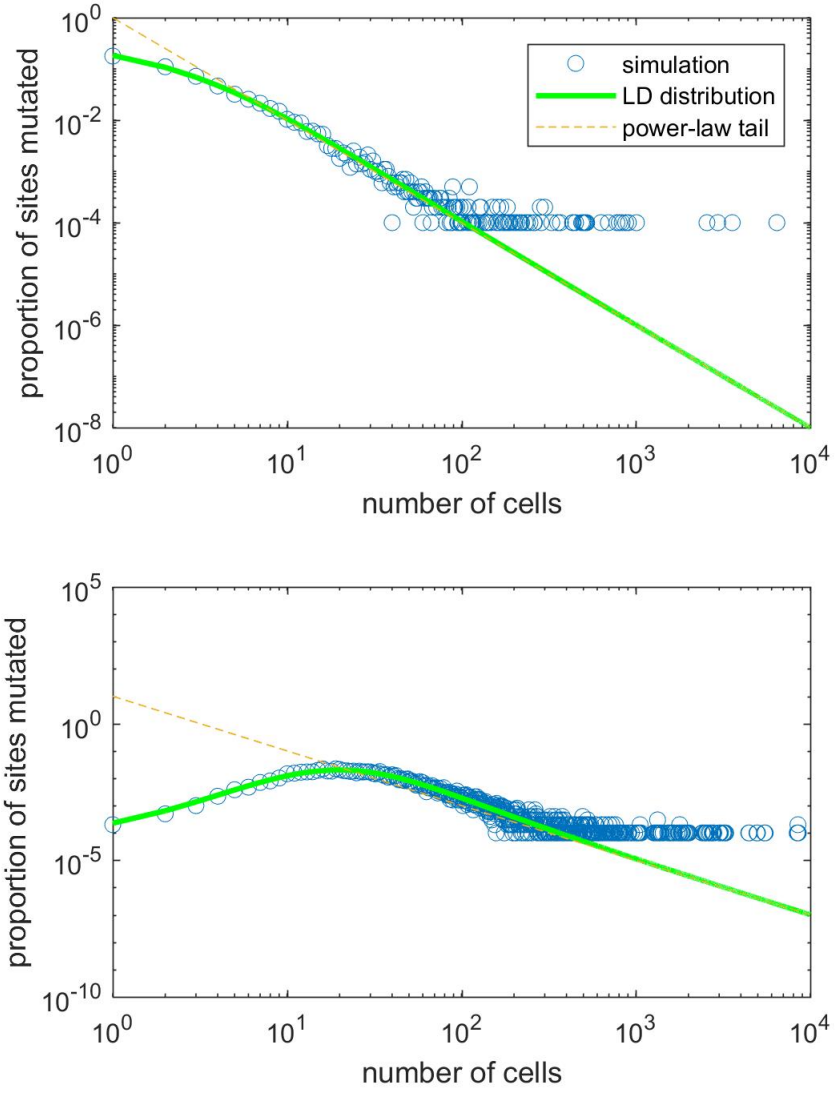


Figure 3.1: The site frequency spectrum is simulated and compared to the Luria-Delbrück (LD) distribution. Each plot is a single realisation of the process. Observe the transition from a deterministic shape at small frequencies to a random point process at large frequencies. The parameters are $\mu = 10^{-4}$ (top) and 10^{-3} (bottom), and $n = |\mathcal{S}| = 10^4$. The parameter values were chosen to be relatively close to 1 for computational ease.

are relatively few cells and few cell divisions. Hence there are few early-arising mutations, which have enough time to reach large frequencies. These mutations may arise on the same branches of the evolutionary tree.

Before discussing dependencies, what happens at a single site? The counterpart to (3.2) is that the proportion of sites which are mutated in at least proportion $\epsilon \in (0,1)$ of cells is

$$\frac{|\{i \in \mathcal{S} : n^{-1}B_i^{n,\mu} \geq \epsilon\}|}{|\mathcal{S}|} \to 0.$$

Similarly, the probability that a single site $i \in \mathcal{S}$ is mutated in at least proportion ϵ of cells is

$$\mathbb{P}[B_i^{n,\mu} \geq \epsilon] \to 0. \tag{3.3}$$

The next result sheds light on (3.3) with greater precision.

Theorem 3.4.1. *Let $(a,b) \subset (0,1)$. As $\mu \to 0$ and $n\mu \to \theta < \infty$,*

$$\mu^{-1}\mathbb{P}[n^{-1}B_i^{n,\mu} \in (a,b)] \to a^{-1} - b^{-1}.$$

Remark 3.4.2. *Theorems 3.3.1 and 3.4.1 are not two isolated approximations. They smoothly connect:*

$$\begin{aligned} \mathbb{P}[n^{-1}B_i^{n,\mu} \in (a,b)] &\approx \mu\left(a^{-1} - b^{-1}\right) && (3.4)\\ &\approx \mathbb{P}[n^{-1}B_i \in (a,b)], && (3.5) \end{aligned}$$

where Approximation (3.4) is Theorem 3.4.1, B_i is the Luria-Delbrück random variable of Theorem 3.3.1, and Approximation (3.5) is the Luria-Delbrück distribution's power-law tail.

Theorem 3.4.1 and Remark 3.4.2 are depicted in Figure 3.2.

According to Theorem 3.4.1 and by linearity of expectation, the expected number of sites which are mutated in a certain proportion of cells is

$$\mathbb{E}\left|\left\{i \in \mathcal{S} : n^{-1}B_i^{n,\mu} \in (a,b)\right\}\right| \approx |\mathcal{S}|\mu\left(a^{-1} - b^{-1}\right). \tag{3.6}$$

This suggests that in order to witness large-frequency mutations it is necessary to take the number of sites at least as large as the mutation rate is small. Taking $|\mathcal{S}| \to \infty$, (3.6) can be rewritten as the following.

Corollary 3.4.3. *Let $(a,b) \subset (0,1)$. As $\mu \to 0$, $n\mu \to \theta < \infty$, and $|\mathcal{S}|\mu \to \eta < \infty$,*

$$\mathbb{E}\left|\left\{i \in \mathcal{S} : n^{-1}B_i^{n,\mu} \in (a,b)\right\}\right| \to \eta(a^{-1} - b^{-1}).$$

In order to understand more than expectations, to see dependencies between sites, the evolutionary tree needs to be introduced. The tree, following standard notation, is the set

$$\mathcal{T} = \cup_{l=0}^{\infty}\{0,1\}^l,$$

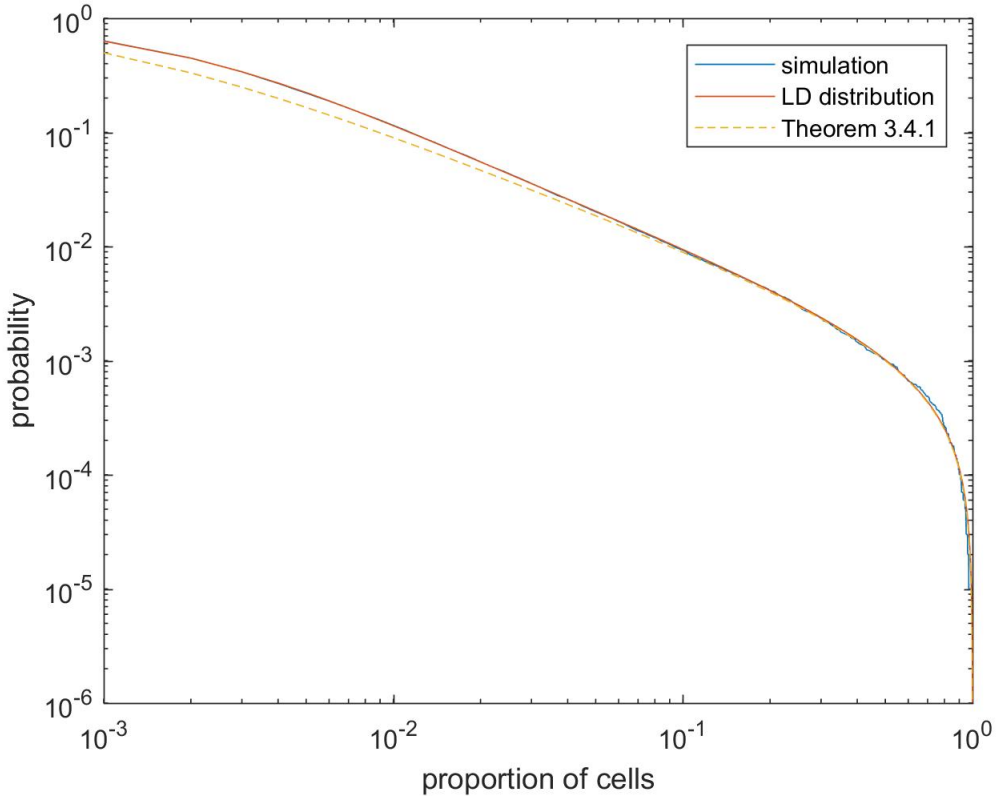


Figure 3.2: The number of mutant cells with respect to a single site is simulated 10^5 times. The plot shows the probability that at least proportion $a \in (0,1)$ of cells are mutant. The parameters are $\mu = 10^{-3}$ and $n = 10^3$. Apparently Theorem 3.4.1 offers a good fit for large frequencies, while the Luria-Delbrück distribution appears perfect for all frequencies.

whose elements are the cells. A partial ordering, $\prec$, is defined on $\mathcal{T}$. For $x, y \in \mathcal{T}$, $x \prec y$ means that cell y is a descendant of cell x. That is, $x \prec y$ if and only if

1. there are $l_1, l_2 \in \mathbb{N}_0$ with $l_1 < l_2$ and $x \in \{0,1\}^{l_1}, y \in \{0,1\}^{l_2}$;
2. the first l_1 entries of y agree with the entries of x.

Note that the empty sequence $\emptyset$ is the single element of $\{0,1\}^0 \subset \mathcal{T}$, and $\emptyset \prec x$ for any $x \in \mathcal{T}\backslash\{\emptyset\}$. This means that $\emptyset$ is the initial cell from which all other cells descend. For further notation, write $x0$ and $x1$ for the daughters of $x \in \mathcal{T}$. Precisely, if $x \in \mathcal{T}$ and $j \in \{0,1\}$, then xj is the element of $\{0,1\}^{l+1}$ whose first l entries are the entries of x and whose last entry is j.

The next result shows the relationship between mutation frequencies and the tree structure; it gives the site frequency spectrum at large frequencies.

Theorem 3.4.4. *Let $\epsilon > 0$. As $\mu \to 0$, $n\mu \to \theta < \infty$, and $|\mathcal{S}|\mu \to \eta < \infty$,*

$$\sum_{i \in \mathcal{S}} \delta_{n^{-1} B_i^{n,\mu}} \to \sum_{x \in \mathcal{T}\backslash\{\emptyset\}} M_x \delta_{P_x}$$

in distribution, on the space of measures on $[\epsilon, 1]$ (equipped with the vague topology). Here:

- *$(M_x)_{x \in \mathcal{T}\backslash\{\emptyset\}}$ is a family of i.i.d. Poisson($\eta/2$) random variables.*
- *For each $x \in \mathcal{T}\backslash\{\emptyset\}$, $P_x = \prod_{\emptyset \prec y \preceq x} U_y$, where:*
  - *the U_y are uniform random variables on $[0,1]$;*
  - *for any $y \in \mathcal{T}$, $U_{y0} + U_{y1} = 1$;*
  - *$(U_{y0})_{y \in \mathcal{T}}$ is an independent family, which is independent of $(M_x)_{x \in \mathcal{T}}$.*

To interpret Theorem 3.4.4, M_x is the number of mutations which arise at the birth of cell x, and P_x is the long-term proportion of cells which have descended from cell x. These quantities and their relationship to the infinite-sites assumption are explained in the Theorem's proof, in Section 3.6.4.

Remark 3.4.5. *Taking the expectation of Theorem 3.4.4's limit, Corollary 3.4.3's limit can be recovered:*

$$\mathbb{E} \sum_{x \in \mathcal{T}\backslash\{\emptyset\}} M_x \delta_{P_x}(a,b) = \eta(a^{-1} - b^{-1}).$$

Remark 3.4.6. *The variance of Theorem 3.4.4's limit is not to be ignored:*

$$Var \sum_{x \in \mathcal{T}\backslash\{\emptyset\}} M_x \delta_{P_x}(a,b) \geq \mathbb{E} \sum_{x \in \mathcal{T}\backslash\{\emptyset\}} M_x \delta_{P_x}(a,b).$$

Remark 3.4.7. *Corollary 3.4.3 recovers the same result seen in [39, 7, 10], whose models required the infinite-sites assumption (as discussed in Section 2.8). The proof of Theorem 3.4.4, in particular, elucidates the reasonableness of the infinite-sites assumption. On the other hand, the results of Section 3.3 detail mutation frequencies to a greater level of precision, and there infinite-sites violations make their presence felt.*

3.5 Generalisations

First the model generalisations are introduced, and then results and conjectures are offered.

3.5.1 Cell death

Cells divide and die as a continuous-time Markov branching process.

Without cell death, it is certain that n cells are reached. With cell death, the population may go extinct before n cells are reached. Therefore, in this generalised setting, the $B_i^{n,\mu}$ denote mutation frequencies when n cells are first reached *conditioned on the event that n cells are reached.*

3.5.2 Selection

A cell's division and death rates are a function of its genome.

It will help to classify different types of genetic site. Partition the sites into *neutral* and *selective* sites:

$$\mathcal{S} = \mathcal{S}_{\text{neut}} \cup \mathcal{S}_{\text{sel}}.$$

Mutations at neutral sites do not affect division nor death rates, whereas mutations at selective sites may affect division or death rates. That is to say, division and death rates can be written as a function of the genome at just the selective sites. For a genome $v \in \mathcal{G}$, write $v' = (v_i)_{i \in \mathcal{S}_{\text{sel}}}$ for its restriction to the selective sites. Say that there exist functions $\alpha, \beta : \mathcal{N}^{\mathcal{S}_{\text{sel}}} \to [0, \infty)$, such that a cell with genome $v \in \mathcal{G}$ has division and death rates $\alpha(v')$ and $\beta(v')$. Assume that the initial cell's genome gives a positive growth rate: $\alpha(u') > \beta(u')$.

Rather vaguely, there are two cases to consider: $\mathcal{S}_{\text{sel}}$ is large, and $\mathcal{S}_{\text{sel}}$ is small. We will discuss the latter case.

3.5.3 Heterogeneous mutation rates

Site i mutates from nucleotide χ to nucleotide ψ at rate $\mu_i^{\chi,\psi}$. Let's state this precisely. Suppose that a cell with genome $v = (v_i)_{i \in \mathcal{S}} \in \mathcal{G}$ divides. Then, denoting its daughters' genomes as $(V_i^1)_{i \in \mathcal{S}}$ and $(V_i^2)_{i \in \mathcal{S}}$:

- The V_i^r are independent over $i \in \mathcal{S}$ and $r \in \{1, 2\}$.
- $V_i^r = \psi$, with probability $\mu_i^{v_i,\psi}/2$ for $\psi \in \mathcal{N} \backslash \{v_i\}$;
  $V_i^r = v_i$ otherwise.

Remark 3.5.1. *Taking $\mu_i^{\chi,\psi} = \mu/3$ for $\chi \neq \psi$ recovers the original model.*

Remark 3.5.2. *As before, the factor of $1/2$ in the mutation probabilities balances that fact that 2 cells are produced at cell division.*

Slightly adapting previous notation, write

$$\mu = \left(\mu_i^{\chi,\psi}\right)_{i \in \mathcal{S}; \chi, \psi \in \mathcal{N}}$$

for the collection of mutation rates and keep the notation $B_i^{n,\mu}$ for mutation frequencies.

3.5.4 Results

To begin, Theorem 3.3.1 is generalised. The genomes whose only difference from the initial cell's genome is at site $i \in \mathcal{S}$,

$$\mathcal{G}_i = \{v \in \mathcal{G} : \forall j \in \mathcal{S}, (u_j \neq v_j \iff i = j)\}, \tag{3.7}$$

will play a crucial role.

Theorem 3.5.3. *Take $\mu_i^{\chi,\psi} \to 0$ and $n\mu_i^{\chi,\psi} \to \theta_i^{\chi,\psi} < \infty$ for all $i \in \mathcal{S}$ and $\chi, \psi \in \mathcal{N}$ with $\chi \neq \psi$. Then*

$$(B_i^{n,\mu})_{i\in\mathcal{S}} \to \left(\sum_{v\in\mathcal{G}_i} X_v\right)_{i\in\mathcal{S}}$$

in distribution, where

- *the X_v are independent;*
- *X_v has Luria-Delbrück distribution with parameters*

$$\left(\frac{\alpha(v')}{\alpha(u') - \beta(u')}, \frac{\beta(v')}{\alpha(u') - \beta(u')}, \theta_i^{u_i,v_i}\right).$$

What about taking the number of sites to infinity? The distinction between neutral and selective sites becomes important. The number of neutral sites will be taken to infinity, while the selective sites remain finite (which is the meaning of '$\mathcal{S}_{\text{sel}}$ is small' in Section 3.5.2). Heterogeneous mutation rates also require attention for the infinite-sites limit. Partition the set of neutral sites:

$$\mathcal{S}_{\text{neut}} = \bigcup_{j\in J} \mathcal{S}(j),$$

such that mutation rates and the initial genome's nucleotides are homogeneous on $\mathcal{S}(j)$ (J is just some indexing set). For $i \in \mathcal{S}(j)$, write $\mu_i^{\chi,\psi} = \mu^{\chi,\psi}(j)$ for the mutation rates and $u_i = u(j)$ for the initial genome's nucleotide. The ratios $|\mathcal{S}(j)|/|\mathcal{S}_{\text{neut}}|$ will be assumed to converge. Now Corollary 3.3.2, which shows a deterministic limit for the site frequency spectrum, is generalised.

Corollary 3.5.4. *Take $\mu^{\chi,\psi}(j) \to 0$, $n\mu^{\chi,\psi}(j) \to \theta^{\chi,\psi}(j) < \infty$, $|\mathcal{S}_{neut}| \to \infty$, and $|\mathcal{S}(j)|/|\mathcal{S}_{neut}| \to q(j)$, for all $j \in J$ and $\chi, \psi \in \mathcal{N}$ with $\chi \neq \psi$. Then*

$$\frac{|\{i \in \mathcal{S} : B_i^{n,\mu} = k\}|}{|\mathcal{S}|} \xrightarrow{p} \sum_{j\in J} q(j)\mathbb{P}[B(j) = k],$$

where

$$B(j) = \sum_{\psi \in \mathcal{N} \setminus \{u(j)\}} X^{\psi}(j),$$

and the $X^{\psi}(j)$ *are independent Luria-Delbrück random variables with parameters*

$$\left(\frac{\alpha(u')}{\alpha(u') - \beta(u')}, \frac{\beta(u')}{\alpha(u') - \beta(u')}, \theta^{u(j),\psi}(j) \right).$$

Selection is visible in Theorem 3.5.3. By contrast, selection is invisible in Corollary 3.5.4. This is because mutations at selective sites do little to change the evolutionary tree's structure, and the vast number of neutral mutations dominate the picture.

3.5.5 Conjectures

Next a generalisation of Corollary 3.4.3 is conjectured.

Conjecture 3.5.5. *Take* $\mu^{\chi,\psi}(j) \to 0$, $n\mu^{\chi,\psi}(j) \to \theta^{\chi,\psi}(j)$, $\mu^{\chi,\psi}(j)|\mathcal{S}(j)| \to \eta^{\chi,\psi}(j)$, *for all* $j \in J$ *and* $\chi, \psi \in \mathcal{N}$ *with* $\chi \neq \psi$. *Then*

$$\begin{aligned} &\mathbb{E} \left| \left\{ i \in \mathcal{S} : n^{-1} B_i^{n,\mu} \in (a,b) \right\} \right| \\ &\quad \to \frac{\alpha(u')}{\alpha(u') - \beta(u')} (a^{-1} - b^{-1}) \sum_{j \in J} \sum_{\psi \in \mathcal{N} \setminus \{u(j)\}} \eta^{u(j),\psi}(j). \end{aligned}$$

Lastly, a generalisation of Theorem 3.4.4 is conjectured.

Conjecture 3.5.6. *Let* $\epsilon > 0$. *Take* $\mu^{\chi,\psi}(j) \to 0$, $n\mu^{\chi,\psi}(j) \to \theta^{\chi,\psi}(j)$, $\mu^{\chi,\psi}(j)|\mathcal{S}(j)| \to \eta^{\chi,\psi}(j)$, *for all* $j \in J$ *and* $\chi, \psi \in \mathcal{N}$ *with* $\chi \neq \psi$. *Then*

$$\sum_{i \in \mathcal{S}} \delta_{n^{-1} B_i^{n,\mu}} \to \sum_{x \in \mathcal{T} \setminus \{\emptyset\}} M_x \delta_{P_x}$$

in distribution, on the space of measures on $[\epsilon, 1]$ *(equipped with the vague topology). Here,*

- $\mathcal{T}$ *and* (P_x) *are unchanged from Theorem 3.4.4;*
- (M_x) *is an i.i.d. family, but for* $\beta > 0$ *it is not independent of* (P_x)*;*
- $$\mathbb{E} M_x = \frac{\alpha(u') + \beta(u')}{2(\alpha(u') - \beta(u'))} \sum_{j \in J} \sum_{\psi \in \mathcal{N} \setminus \{u(j)\}} \eta^{u(j),\psi}(j).$$

Conjectures 3.5.5 and 3.5.6 are explained and heuristically derived in Section 3.7.

Theorem 3.5.3, Corollary 3.5.4, and Conjectures 3.5.5 and 3.5.6 all suggest a biologically pertinent statement: selection has no impact on most neutral mutations. The caveat is that selective sites are assumed to be few. Consider the opposite, that selective sites are many. Then, with non-negligible probability,

selective sites mutate early in the growth trajectory on the same branches of the evolutionary tree. Hence there is a highly complex interplay between the tree structure and genetic information, which is beyond the scope of the thesis. Questions are open for future research.

Detecting selection in cancer through the lens of mutation frequencies is foggy territory. How or whether it is even possible to do so is a subject of current debate in the biological literature. See for example [38].

3.6 Proofs

3.6.1 Theorems 3.3.1 and 3.5.3

A more detailed result

Theorem 3.3.1 and its generalisation, Theorem 3.5.3, can be understood in the light of a more detailed result. First some notation needs to be introduced. Write

$$X_v^\mu(t) \tag{3.8}$$

for the number of cells with genome $v \in \mathcal{G}$ at time $t \geq 0$. Write

$$\sigma_n^\mu = \min\left\{t \geq 0 : \sum_{v \in \mathcal{G}} X_v^\mu(t) = n\right\}$$

for the time at which n cells are reached. Write

$$((X_v^\mu(\sigma_n^\mu))_{v\in\mathcal{G}} | \sigma_n^\mu < \infty)$$

for the number of cells of each genotype at time σ_n^μ, conditioned on $\sigma_n^\mu < \infty$ (i.e. the event that n cells are reached). Recall from (3.7) that $\mathcal{G}_i$ is the subset of genomes with exactly one mutation which is at site i. Write

$$\mathcal{G}_{\geq 2} = \{v \in \mathcal{G} : |\{i \in \mathcal{S} : v_i \neq u_i\}| \geq 2\}$$

for the subset of genomes with at least two mutations.

Theorem 3.6.1. *For $i \in \mathcal{S}$ and $\chi, \psi \in \mathcal{N}$ with $\chi \neq \psi$, take $\mu_i^{\chi,\psi} \to 0$ and $n\mu_i^{\chi,\psi} \to \theta_i^{\chi,\psi} \in (0, \infty)$. Then*

$$((X_v^\mu(\sigma_n^\mu))_{v\in\mathcal{G}} | \sigma_n^\mu < \infty) \to (X_v)_{v\in\mathcal{G}}$$

in distribution, where

- *the X_v are independent;*
- *for $v \in \mathcal{G}_i$, X_v has Luria-Delbrück distribution with parameters*

$$\left(\frac{\alpha(v')}{\alpha(u') - \beta(u')}, \frac{\beta(v')}{\alpha(u') - \beta(u')}, \theta_i^{u_i,v_i}\right);$$

- $X_u = \infty$;
- *for* $v \in \mathcal{G}_{\geq 2}$, $X_v = 0$.

Remark 3.6.2. *Theorems 3.3.1 and 3.5.3 are a consequence of Theorem 3.6.1 via the continuous mapping theorem, because*

$$B_i^{n,\mu} = \left(\sum_{\substack{v \in \mathcal{G} \\ v_i \neq u_i}} X_v^\mu(\sigma_n^\mu) \middle| \sigma_n^\mu < \infty \right).$$

We break down the proof of Theorem 3.6.1 into four parts. First, we present a construction of $(X_v^\mu(\sigma_n^\mu))_{v\in\mathcal{G}}$. The construction will illuminate the importance of various subpopulations and the mutations between them. Second, we show that cell divisions which produce two mutant daughter cells can be neglected. Third, we show that some mutation times converge to a Poisson process and therefore $(X_v^\mu(\sigma_n^\mu))_{v\in\mathcal{G}}$ converges (conditioned on a certain event). Fourth, we condition on $\{\sigma_n^\mu < \infty\}$ to conclude the proof.

Additional notation to be used in the proof: for $v \in \mathcal{G}$,

$$e_v = (\delta_{v,w})_{w\in\mathcal{G}}$$

is the element of $(\mathbb{N}_0)^{\mathcal{G}}$ denoting that there is one genome v and zero other genomes.

Construction

Let

$$[\mu_n]_{n\in\mathbb{N}} = \left[\left(\mu_{n,i}^{\chi,\psi} \right)_{i\in\mathcal{S}:\chi,\psi\in\mathcal{N}} \right]_{n\in\mathbb{N}}$$

be a sequence of mutation rates with

$$\lim_{n\to\infty} n\mu_{n,i}^{\chi,\psi} = \theta_i^{\chi,\psi} < \infty$$

for $\chi \neq \psi$. Fix $n \in \mathbb{N}$. For $v, w \in \mathcal{G}$, write

$$p_n(v,w) = \prod_{i\in\mathcal{S}} \mu_{n,i}^{u_i,v_i} \mu_{n,i}^{u_i,w_i}/4 \tag{3.9}$$

for the probability that a cell with genome u which divides, gives daughters with genomes v, w (the daughter cells are distinguished - so v, w is different to w, v). Now the construction of $(X_v^{\mu_n}(\sigma_n^{\mu_n}))_{v\in\mathcal{G}}$ begins. For the foundational step, introduce the following random variables on a fresh probability space.

1. $$(Z^n(t))_{t\geq 0}$$

   is a birth-death branching process with birth and death rates

   $$\alpha_n := \alpha(u')p_n(u,u)$$

and

$$\beta_n := \beta(u') + \alpha(u') \sum_{v,w \in \mathcal{G}\backslash\{u\}} p_n(v,w).$$

The initial condition $Z^n(0) = 1$ is assumed.

2. For $j \in \mathbb{N}$,

$$E_j^n$$

are $\{\emptyset\} \cup (\mathcal{G}\backslash\{u\})^2$-valued random variables, with

$$\mathbb{P}[E_j^n = \emptyset] = \frac{\beta(u')}{\beta_n},$$

and for $v, w \in \mathcal{G}\backslash\{u\}$

$$\mathbb{P}[E_j^n = (v,w)] = \frac{\alpha(u')p_n(v,w)}{\beta_n}.$$

3. For $v \in \mathcal{G}\backslash\{u\}$ and $j \in \mathbb{N}$,

$$\mathcal{Y}_{v,j}^n(\cdot)$$

is distributed as $(X_x^{\mu_n}(\cdot))_{x\in\mathcal{G}}$ (defined in (3.8)) but with initial condition $\mathcal{Y}_{v,j}^n(0) = e_v$.

4. For $v, w \in \mathcal{G}\backslash\{u\}$ and $j \in \mathbb{N}$,

$$\mathcal{Y}_{v,w,j}^n(\cdot)$$

is distributed as $(X_x^{\mu_n}(\cdot))_{x\in\mathcal{G}}$ but with initial condition $\mathcal{Y}_{v,w,j}^n(0) = e_v + e_w$.

5. For $v \in \mathcal{G}$,

$$(N_v(t))_{t\geq 0}$$

are Poisson counting processes with rate 1.

The random variables

$$\left[Z^n(\cdot), E_j^n, \mathcal{Y}_{v,j}^n(\cdot), \mathcal{Y}_{v,w,j}^n(\cdot), N_v(\cdot)\right] \tag{3.10}$$

are assumed to be independent ranging over v, w, j.

Let's explain the meaning of the random variables introduced so far. $Z^n(\cdot)$ represents the 'primary' subpopulation - which we define as the type u cells whose ancestors are all of type u. That is to say, there is an unbroken lineage of type u cells between any primary cell and the initial cell. The rate, α_n, that a primary cell gives birth to another primary cell, is simply the type u division rate multiplied by the probability that no mutation occurs in either daughter cell. The rate, β_n, that a primary cell is removed, is the rate that a type u cell dies plus the rate that a type u cell divides to produce two mutant daughter cells.

The E_j^n describe what happens at the jth downstep in the primary subpopulation trajectory. If $E_j^n = \emptyset$, then the downstep is a primary cell death. If $E_j^n = (v,w)$, then the downstep is a primary cell dividing to produce two mutant daughter cells of types v and w.

Sometimes a primary cell divides to produce one primary cell and one mutant cell of type v. For the jth time that this occurs, $\mathcal{Y}^n_{v,j}(t)$ is the vector which counts the cells of each genotype amongst the descendants of that type v cell, time t after its birth.

Sometimes a primary cell divides to produce two mutant cells of types v and w. For the jth time that this occurs, $\mathcal{Y}^n_{v,w,j}(t)$ is the vector which counts the cells of each genotype amongst the descendants of the two mutants time t after their birth.

The $N_v(\cdot)$ will soon be rescaled in time to represent the times at which primary cells divide to produce one primary cell and one cell with genome v.

The random variables introduced so far, seen together in (3.10), provide all the neccessary ingredients for the construction of $(X^{\mu_n}_v(\sigma^{\mu_n}_n))_{v\in\mathcal{G}}$. Now we build upon these founding objects, defining further random variables.

6. For $v \in \mathcal{G}\backslash\{u\}$ and $t \geq 0$,

$$K^n_v(t) = N_v\left(2p_n(u,v)\alpha(u')\int_0^t Z^n(s)ds\right). \tag{3.11}$$

7. For $j \in \mathbb{N}$ and $v \in \mathcal{G}\backslash\{u\}$,

$$T^n_{v,j} = \min\{t \geq 0 : K^n_v(t) = j\}.$$

8. $$S^n_1 = \min\{t \geq 0 : Z^n(t) - Z^n(t^-) = -1\},$$

   and then for $j > 1$, recursively,

$$S^n_j = \min\{t > S^n_{j-1} : Z^n(t) - Z^n(t^-) = -1\}.$$

   (Here $Z^n(t^-) := lim_{s\uparrow t} Z^n(s)$.)

9. For $v, w \in \mathcal{G}\backslash\{u\}$,

$$T^n_{v,w,1} = \min\{S^n_j : j \in \mathbb{N}, E^n_j = (v,w)\},$$

   and then for $j > 1$, recursively,

$$T^n_{v,w,j} = \min\{S^n_j : j \in \mathbb{N}, S^n_j > T^n_{v,w,j-1}, E^n_j = (v,w)\}.$$

10. For $v, w \in \mathcal{G}\backslash\{u\}$, and $t \geq 0$,

$$K^n_{v,w}(t) = \#\{j \in \mathbb{N} : T^n_{v,w,j} \leq t\}.$$

Let's explain the meaning of the new random variables. The $K^n_v(t)$ specify the number of times before time t that primary cells have divided to produce one primary cell and one type v cell. Let's check that this interpretation makes sense. Conditioned on the trajectory of $Z^n(\cdot)$, $K^n_v(\cdot)$ is certainly a Markov process, and increases by 1 at rate $2p_n(u,v)\alpha(u')Z^n(t)$ - i.e. the rate at which primary cells

divide multiplied by the probability that exactly one daughter cell is primary and one is type v.

S_j^n is the time of the jth downstep of the primary subpopulation size. Then $T_{v,w,j}^n$ is the time of the jth primary cell division which produces cells of types v and w. $K_{v,w}^n(t)$ is the number of primary cell divisions before time t which produce cells of types v and w.

At last the construction reaches its denouement.

11. For $t \geq 0$,

$$\begin{aligned}\mathcal{X}^n(t) &= Z^n(t)e_u \\ &\quad + \sum_{v \in \mathcal{G}\setminus\{u\}} \sum_{j=1}^{K_v^n(t)} \mathcal{Y}_{v,j}^n(t - T_{v,j}^n) \\ &\quad + \sum_{v,w \in \mathcal{G}\setminus\{u\}} \sum_{j=1}^{K_{v,w}^n(t)} \mathcal{Y}_{v,w,j}^n(t - T_{v,w,j}^n).\end{aligned}$$

12.
$$\sigma_n = \min\{t \geq 0 : |\mathcal{X}^n(t)| = n\},$$
where $|\cdot|$ is the l_1-norm on $(\mathbb{N}_0)^{\mathcal{G}}$.

Upon reflection it should be clear that $\mathcal{X}^n(\cdot)$ has the same distribution as $(X_v^{\mu_n}(\cdot))_{v\in\mathcal{G}}$. Both objects are Markov processes on $(\mathbb{N}_0)^{\mathcal{G}}$, whose initial conditions and transition rates coincide. Next we show that certain elements of the construction converge in distribution.

Lemma 3.6.3. *As $n \to \infty$,*

$$(e^{-\lambda_n t} Z^n(t))_{t\in[0,\infty]} \to (e^{-\lambda t} Z^*(t))_{t\in[0,\infty]}$$

in distribution, on the space $\mathbb{D}([0,\infty],\mathbb{R})$. Here

$$\lambda_n := \alpha_n - \beta_n$$

is the growth rate of the primary cell population,

$$\lambda := \alpha(u') - \beta(u')$$

is the large n limit of λ_n, and $Z^*(\cdot)$ is a birth-death branching process with birth and death rates $\alpha(u')$ and $\beta(u')$.

Remark 3.6.4. *The processes of Lemma 3.6.3 are defined on the timeline extended to include infinity. For n large enough that $\lambda_n > 0$,*

$$e^{-\lambda_n \infty} Z^n(\infty) := \lim_{t\to\infty} e^{-\lambda_n t} Z^n(t) = W^n,$$

and

$$e^{-\lambda \infty} Z^*(\infty) := \lim_{t\to\infty} e^{-\lambda t} Z^*(t) = W^*.$$

The limits W^n and W^* exist and are finite almost surely (seen in Lemma 1.2.3).

Proof of Lemma 3.6.3. Convergence in finite dimensional distributions is immediate. To show tightness we shall use Aldous's criterion [1]. His result is for the time interval $[0,1]$; so let's identify $[0,1]$ with $[0,\infty]$, by $t(s) = -\lambda_n^{-1}\log(1-s)$ for $s \in [0,1]$. We check tightness of the sequence of martingales $(M^n(s))_{s\in[0,1]}$ defined by

$$\begin{aligned} M^n(s) &:= e^{-\lambda_n t(s)} Z^n(t(s)) \\ &= (1-s)Z^n(-\lambda_n^{-1}\log(1-s)). \end{aligned}$$

Suppose that (ρ_n) is a sequence of stopping times with respect to $(M^n(\cdot))$, and (δ_n) is a positive sequence converging to zero. Then, writing $\mathcal{F}_{\rho_n}$ for the sigma-algebra generated by $M^n(\cdot)$ up to time ρ_n,

$$\begin{aligned} \mathbb{E}[(M^n(\rho_n+\delta_n) - M^n(\rho_n))^2 | \mathcal{F}_{\rho_n}] &= \mathbb{E}[M^n(\rho_n+\delta_n)^2|\mathcal{F}_{\rho_n}] - M^n(\rho_n)^2 \\ &= \delta_n \frac{\alpha_n+\beta_n}{\lambda_n} M^n(\rho_n), \end{aligned}$$

where the last equality comes thanks to the fact that

$$M^n(s)^2 - \frac{\alpha_n+\beta_n}{\lambda_n}(1-s)M^n(s)$$

is a martingale. Now,

$$\begin{aligned} \mathbb{E}[(M^n(\rho_n+\delta_n) - M^n(\rho_n))^2] &= \delta_n \frac{\alpha_n+\beta_n}{\lambda_n} \mathbb{E}M^n(\rho_n) \\ &= \delta_n \frac{\alpha_n+\beta_n}{\lambda_n}. \end{aligned}$$

Take $n \to \infty$ to see that $M^n(\rho_n+\delta_n) - M^n(\rho_n)$ converges to zero in L_2 and hence in probability, thus satisfying Aldous's criterion. □

Lemma 3.6.5. *As $n \to \infty$,*

$$\mathcal{Y}^n_{v,j}(\cdot) \to Y_{v,j}(\cdot)e_v$$

in distribution, where $Y_{v,j}(\cdot)$ is a birth-death branching process with birth and death rates $\alpha(v')$ and $\beta(v')$ and initial condition $Y_{v,j}(0) = 1$. The convergence is in the space $\mathbb{D}([0,\infty), \mathbb{R}^{\mathcal{G}})$.

Proof. It is enough to note that the transition rates converge (see for example page 262 of [14]). □

Lemma 3.6.6. *As $n \to \infty$,*

$$\left(\sum_{j \le kn^{3/2}} 1_{\{E^n_j \neq \emptyset\}} \right)_{k\in\mathbb{N}} \to (0)$$

in distribution, on the space $\mathbb{R}^{\mathbb{N}}$.

Proof. Note that

$$\lim_{n\to\infty} n^{3/2}\mathbb{P}[E_j^n \neq \emptyset] = 0.$$

Then

$$\begin{aligned}\mathbb{P}\left[\sum_{j\leq kn^{3/2}} 1_{\{E_j^n\neq\emptyset\}} = 0; k = 1,..,r\right] &= \mathbb{P}\left[E_j^n = \emptyset; j \leq rn^{3/2}\right] \\ &= \left(1 - \mathbb{P}\left[E_j^n \neq \emptyset\right]\right)^{\lfloor rn^{3/2}\rfloor} \\ &\to 1.\end{aligned}$$

□

Remark 3.6.7. *The number* 3/2 *which appears in Lemma 3.6.6 is not special. It only matters that* $3/2 \in (1,2)$. *The relevance of the result will be seen in Section 3.6.1.*

We are yet to say how the random variables in (3.10) are jointly distributed over $n \in \mathbb{N}$. In fact, the choice of this joint distribution over $n \in \mathbb{N}$ has no relevance to the statement of Theorem 3.6.1. Hence the choice can be freely made, in a way that streamlines the proof. We assume that:

$$\lim_{n\to\infty} (e^{-\lambda_n t} Z^n(t))_{t\in[0,\infty]} = (e^{-\lambda t} Z^*(t))_{t\in[0,\infty]} \tag{3.12}$$

almost surely, on the space $\mathbb{D}([0,\infty],\mathbb{R})$;

$$\lim_{n\to\infty} (\mathcal{Y}_{v,j}^n(t))_{t\in[0,\infty)} = (Y_{v,j}(t)e_v)_{t\in[0,\infty)} \tag{3.13}$$

almost surely, on the space $\mathbb{D}([0,\infty),\mathbb{R}^{\mathcal{G}})$, for $v \in \mathcal{G}\backslash\{u\}$ and $j \in \mathbb{N}$; and

$$\left(\sum_{j\leq kn^{3/2}} 1_{\{E_j^n\neq\emptyset\}}\right)_{k\in\mathbb{N}} \to (0) \tag{3.14}$$

almost surely, on the space $\mathbb{R}^{\mathbb{N}}$.

To justify that it is possible to have constructed the random variables in such a way that (3.12), (3.13), and (3.14) hold, one can bring in Skorokhod's Representation Theorem, to use with the Lemmas 3.6.3, 3.6.5, and 3.6.6.

Double mutations can be neglected

Call the event that a primary cell divides to produce two mutant cells a 'double mutation'. Recall that double mutations are represented by the events $\{E_j^n = (v,w)\}$, which occur at the times S_j^n when the primary cell population steps down in size. In order to comment on double mutations, we will first prove a rather crude upper bound for the number of downsteps in the primary cell population

trajectory. Write

$$\tau_n := \min\{t \geq 0 : Z^n(t) \in \{0, n\}\}, \tag{3.15}$$

for the time at which the primary cell population hits 0 or n. Write

$$D_n := \left|\{j \in \mathbb{N} : S_j^n \leq \tau_n\}\right|$$

for the number of downsteps in the primary cell population before time τ_n.

Lemma 3.6.8.

$$\sup_{n\in\mathbb{N}} n^{-3/2} D_n < \infty$$

almost surely.

Proof. For each $n \in \mathbb{N}$, let $(R_j^n)_{j\in\mathbb{N}}$ be a sequence of i.i.d. random variables with

$$\mathbb{P}[R_j^n = x] = \begin{cases} \alpha_n/(\alpha_n + \beta_n), & x = 1; \\ \beta_n/(\alpha_n + \beta_n), & x = -1; \end{cases}$$

so

$$\left(1 + \sum_{j=1}^{k} R_j^n\right)_{k\in\mathbb{N}}$$

is a random walk, whose distribution matches the steps of $Z^n(\cdot)$. Write

$$\rho_n = \min\left\{k \in \mathbb{N} : 1 + \sum_{j=1}^{k} R_j^n \in \{0, n\}\right\}$$

for the number of steps until the walk hits n or 0. Then the number of downsteps before hitting n or 0 is

$$D_n \stackrel{d}{=} \sum_{j=1}^{\rho_n} 1_{\{R_j^n = -1\}} \leq \rho_n.$$

Therefore we can bound the tail of D_n's distribution:

$$\mathbb{P}[D_n > n^{3/2}] \leq \mathbb{P}[\rho_n > n^{3/2}].$$

But $\{\rho_n > n^{3/2}\} \subset \{1 + \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n < n\}$, so

$$\begin{aligned}
\mathbb{P}[D_n > n^{3/2}] &\leq \mathbb{P}\left[1 + \sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n < n\right] \\
&\leq \mathbb{P}\left[\left(\sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n - \lfloor n^{3/2} \rfloor \lambda_n\right)^2 > \left(\lfloor n^{3/2} \rfloor \lambda_n + 1 - n\right)^2\right] \quad (3.16) \\
&\leq \left(\lfloor n^{3/2} \rfloor \lambda_n + 1 - n\right)^{-2} \operatorname{Var}\left[\sum_{j=1}^{\lfloor n^{3/2} \rfloor} R_j^n\right] \quad (3.17) \\
&\leq cn^{-3/2}, \quad (3.18)
\end{aligned}$$

for some constant $c > 0$. Inequality (3.16) holds for large enough n and Inequality (3.17) is Chebyshev's inequality. Finally, (3.18) gives that

$$\sum_{n \in \mathbb{N}} \mathbb{P}[D_n > n^{3/2}] < \infty,$$

and the result is proven by Borel-Cantelli. □

Now it is to be seen that double mutations occurring before time τ_n can be neglected.

Lemma 3.6.9. *Let $v, w \in \mathcal{G} \backslash \{u\}$. As $n \to \infty$,*

$$K_{v,w}^n(\tau_n) \to 0$$

almost surely.

Proof. From Lemma 3.6.8, $C := \sup_{n \in \mathbb{N}} n^{-3/2} D_n < \infty$. Then

$$\begin{aligned}
K_{v,w}^n(\tau_n) &= \sum_{j=1}^{D_n} 1_{\{E_j^n = (v,w)\}} \\
&\leq \sum_{j=1}^{\lceil Cn^{3/2} \rceil} 1_{\{E_j^n = (v,w)\}} \\
&\leq \sum_{j=1}^{\lceil Cn^{3/2} \rceil} 1_{\{E_j^n \neq \emptyset\}}.
\end{aligned}$$

By (3.14) this converges to zero as $n \to \infty$. □

Convergence

Our next task is to show that $\mathfrak{X}^n(\sigma_n)$ converges when conditioned on the event $\{W^* > 0\}$ (W^* is defined in Remark 3.6.4). The times τ_n (defined in (3.15)) will play the role of a helpful stepping stone in the proof.

Lemma 3.6.10. *Condition on* $\{W^* > 0\}$. *Then, almost surely,*

1. *there exists* n_0 *such that for all* $n \geq n_0$, $Z^n(\tau_n) = n$; *and*
2. $\lim_{n\to\infty} \tau_n = \infty$.

Proof. To see the first statement, observe that there exists n_0 such that for all $n \geq n_0$, $W^n > W^*/2 > 0$. For such n, $\lim_{t\to\infty} Z^n(t) = \infty$, and hence $Z^n(\cdot) > 0$. To see the second statement, suppose for a contradiction that there exists a bounded subsequence $(\tau_{n_k}) \subset [0, C]$. Then, for large enough k,

$$n_k = Z^{n_k}(\tau_{n_k}) \leq \sup_{n\in\mathbb{N}} \sup_{t\in[0,C]} Z^n(t).$$

The left hand side of the inequality is unbounded over k. On the other hand, the right hand side, which does not depend on k, is finite thanks to (3.12). □

Lemma 3.6.11. *Condition on* $\{W^* > 0\}$. *Suppose that* (a_n) *is a real sequence which converges to infinity, with*

$$a_n \leq \tau_n$$

and

$$\lim_{n\to\infty} (a_n - \tau_n) = l \in [-\infty, 0]$$

almost surely. Then for each $t \in \mathbb{R}$, *almost surely,*

$$K_v^n(a_n + t) \to \begin{cases} K_v^*(l + t), & v \in \mathcal{G}_i; \\ 0, & v \in \mathcal{G}_{\geq 2}; \end{cases}$$

where

$$K_v^*(s) = N_v(\lambda^{-1}\alpha(v')\theta_i^{u_i,v_i} e^{\lambda s}).$$

Proof. Thanks to (3.12):

1. for any sequence (t_n) which converges to infinity, $\lim_{n\to\infty} e^{-\lambda_n t_n} Z^n(t_n) = W^*$; and
2. $\sup_{n\in\mathbb{N}} \sup_{t\in[0,\infty]} e^{-\lambda_n t} Z^n(t) < \infty$.

Therefore, for $t \in \mathbb{R}$,

$$\begin{aligned} n^{-1} \int_0^{a_n+t} Z^n(s)ds &= \int_{-a_n}^{t} \frac{Z^n(a_n + s)}{e^{\lambda_n(a_n+s)}} \frac{e^{\lambda_n\tau_n}}{Z^n(\tau_n)} e^{\lambda_n(a_n-\tau_n+s)} ds \\ &\to \int_{-\infty}^{t} e^{\lambda(l+s)} ds \\ &= \lambda^{-1} e^{\lambda(l+t)} \end{aligned}$$

as $n \to \infty$, by dominated convergence. Note also that

$$\lim_{n\to\infty} np_n(u, v) = \begin{cases} \theta_i^{u_i,v_i}/2, & v \in \mathcal{G}_i; \\ 0, & v \in \mathcal{G}_{\geq 2}. \end{cases}$$

Then the result follows straight from the definition of $K_v^n(\cdot)$ in (3.11). □

Lemma 3.6.12. *Condition on* $\{W^* > 0\}$. *Suppose that* (a_n) *satisfies the conditions of Lemma 3.6.11. Then, almost surely,*

$$\lim_{n\to\infty} (\mathcal{X}^n(a_n) - Z^n(a_n)e_u) = \sum_{i\in\mathcal{S}} \sum_{v\in\mathcal{G}_i} e_v \sum_{j=1}^{K_v^*(l)} Y_{v,j}(l - T_{v,j}^*),$$

where $K_v^*(\cdot)$ *is defined in Lemma 3.6.11 and* $T_{v,j}^* = \min\{t \in \mathbb{R} : K_v^*(t) = j\}$.

Proof. Recall that

$$\begin{aligned} \mathcal{X}^n(a_n) - Z^n(a_n)e_u \;=\; & \sum_{v\in\mathcal{G}\setminus\{u\}} \sum_{j=1}^{K_v^n(a_n)} \mathcal{Y}_{v,j}^n(a_n - T_{v,j}^n) \\ & + \sum_{v,w\in\mathcal{G}\setminus\{u\}} \sum_{j=1}^{K_{v,w}^n(a_n)} \mathcal{Y}_{v,w,j}^n(a_n - T_{v,w,j}^n). \end{aligned} \tag{3.19}$$

The 'double mutation' term in (3.19) converges to zero, because

$$K_{v,w}^n(a_n) \le K_{v,w}^n(\tau_n),$$

which converges to zero by Lemma 3.6.9. The 'single mutation' term in (3.19) converges to the required limit due to Lemma 3.6.11 and (3.13). □

Lemma 3.6.13. *Condition on* $\{W^* > 0\}$.

$$\lim_{n\to\infty} (\sigma_n - \tau_n) = 0$$

almost surely.

Proof. By Lemma 3.6.10, for large enough n, $Z^n(\tau_n) = n$. So $|\mathcal{X}^n(\tau_n)| \ge n$, and hence $\sigma_n \le \tau_n$. Therefore

$$\liminf_{n\to\infty} (\sigma_n - \tau_n) \le 0.$$

Suppose, looking for a contradiction, that

$$\liminf_{n\to\infty} (\sigma_n - \tau_n) = l \in [-\infty, 0).$$

Take a subsequence with

$$\lim_{k\to\infty} (\sigma_{n_k} - \tau_{n_k}) = l.$$

Then, by Lemma 3.6.12,

$$|\mathcal{X}^n(\sigma_{n_k}) - Z^n(\sigma_{n_k})e_u|$$

converges, and so must be a bounded sequence. However it is also true that,

taking $k \to \infty$,

$$\begin{aligned} |\mathcal{X}^n(\sigma_{n_k}) - Z^n(\sigma_{n_k})e_u| &= n_k - Z^{n_k}(\sigma_{n_k}) \\ &= n_k \left(1 - \frac{Z^{n_k}(\sigma_{n_k})}{e^{\lambda_{n_k}\sigma_{n_k}}} \frac{e^{\lambda_{n_k}\tau_{n_k}}}{Z^{n_k}(\tau_{n_k})} e^{\lambda_{n_k}(\sigma_{n_k}-\tau_{n_k})}\right) \\ &\sim n_k(1 - e^{\lambda l}), \end{aligned}$$

which is unbounded. □

Lemma 3.6.14. *Condition on* $\{W^* > 0\}$.

$$\lim_{n\to\infty} (\mathcal{X}^n(\sigma_n) - Z^n(\sigma_n)e_u) = \sum_{i\in\mathcal{S}} \sum_{v\in\mathcal{G}_i} e_v \sum_{j=1}^{K_v^*(0)} Y_{v,j}(-T_{v,j}^*),$$

almost surely.

Proof. Combine Lemmas 3.6.12 and 3.6.13. □

The limit of Lemma 3.6.14 is a vector of independent Luria-Delbrück distributions:

$$\sum_{i\in\mathcal{S}} \sum_{v\in\mathcal{G}_i} e_v \sum_{j=1}^{K_v^*(0)} Y_{v,j}(-T_{v,j}^*) \stackrel{d}{=} (X_v)_{v\in\mathcal{G}\setminus\{u\}},$$

where the X_v are as stated in Theorem 3.6.1. To complete the proof of Theorem 3.6.1 we need to show that conditioning on $\{W^* > 0\}$ can be translated to conditioning on $\{\sigma_n < \infty\}$.

Conditioning

In order to connect $\{W^* > 0\}$ and $\{\sigma_n < \infty\}$, the next result is the key. It states that these events are approximately the same for large n.

Proposition 3.6.15.

1. $\lim_{n\to\infty} \mathbb{P}[W^* > 0, \sigma_n = \infty] = 0$, *and*
2. $\lim_{n\to\infty} \mathbb{P}[W^* = 0, \sigma_n < \infty] = 0$.

Let's break the proof of Proposition 3.6.15 into several lemmas; the idea is that the random variable W^n be used as an intermediary.

Lemma 3.6.16.

$$\lim_{n\to\infty} \mathbb{P}[W^* > 0, W^n = 0] = 0.$$

Proof. If $W^* > 0$, then there exists n_0, such that for all $n \geq n_0$

$$W^n > \frac{W^*}{2}.$$

So

$$\lim_{n\to\infty} 1_{\{W^*>0, W^n=0\}} = 0.$$

Therefore, by dominated convergence,

$$\mathbb{P}[W^* > 0, W^n = 0] = \mathbb{E}1_{\{W^*>0, W^n=0\}} \to 0.$$

□

Lemma 3.6.17.

$$\mathbb{P}[W^n > 0, \sigma_n = \infty] = 0.$$

Proof. If $W^n > 0$, then $\lim_{t\to\infty} X^n(t) = \infty$, and so $\sigma_n < \infty$. □

Proof of Part 1 of Proposition 3.6.15.

$$\begin{aligned}\mathbb{P}[W^* > 0, \sigma_n = \infty] &= \mathbb{P}[W^* > 0, \sigma_n = \infty, W^n = 0] \\ &\quad +\mathbb{P}[W^* > 0, \sigma_n = \infty, W^n > 0] \\ &\leq \mathbb{P}[W^* > 0, W^n = 0] \\ &\quad +\mathbb{P}[\sigma_n = \infty, W^n > 0] \to 0\end{aligned}$$

as $n \to \infty$, by Lemmas 3.6.16 and 3.6.17. □

The structure for the proof of Part 2 of Proposition 3.6.15 is much the same as that of Part 1. However the details will require a little extra work.

Lemma 3.6.18.

$$\lim_{n\to\infty} \mathbb{P}[W^* = 0, W^n > 0].$$

Proof. Let $\epsilon > 0$. If $W^* = 0$, then there exists n_0 such that for all $n \geq n_0$

$$W^n < \epsilon.$$

So

$$\lim_{n\to\infty} 1_{\{W^*=0, W^n\geq\epsilon\}} = 0.$$

Note that

$$\begin{aligned}1_{\{W^*=0, W^n>0\}} &= 1_{\{W^*=0, W^n\in(0,\epsilon)\}} + 1_{\{W^*=0, W^n\geq\epsilon\}} \\ &\leq 1_{\{W^n\in(0,\epsilon)\}} + 1_{\{W^*=0, W^n\geq\epsilon\}},\end{aligned}$$

and that

$$\sup_{n\in\mathbb{N}} \mathbb{P}[W^n \in (0, \epsilon)] \leq C\epsilon,$$

for some $C > 0$. Now, using dominated convergence,

$$\limsup_{n\to\infty} \mathbb{P}[W^* = 0, W^n > 0] \leq C\epsilon.$$

But $\epsilon > 0$ was arbitrary, giving the result. □

Lemma 3.6.19.

$$\lim_{n\to\infty} \mathbb{P}[W^n = 0, \sigma_n < \infty] = 0.$$

Proof. If the primary population size never reaches n and there are never any mutations, then the total population size never reaches n. That is, if $Z^n(\tau_n) = 0$, $K_v^n(\cdot) = 0$ and $K_{v,w}^n(\cdot) = 0$ for all $v, w \in \mathcal{G}\backslash\{u\}$, then

$$\sup_{t\geq 0} |\mathcal{X}^n(t)| < n,$$

which means that $\sigma_n = \infty$. Equivalently,

$$\begin{aligned}\{\sigma_n < \infty\} &\subset \{Z^n(\tau_n) = n\} \cup \{\exists v, K_v^n(\cdot) \neq 0\} \cup \{\exists (v,w), K_{v,w}^n(\cdot) \neq 0\} \\ &= \{Z^n(\tau_n) = n\} \cup \{\exists v, K_v^n(\cdot) \neq 0\} \\ &\quad \cup \{\exists (v,w), K_{v,w}^n(\cdot) \neq 0, Z^n(\tau_n) = 0\}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}[W^n = 0, \sigma_n < \infty] &\leq \mathbb{P}[W^n = 0 | Z^n(\tau_n) = n] \\ &\quad + \sum_{v\in\mathcal{G}\backslash\{u\}} \mathbb{P}[K_v^n(\cdot) \neq 0 | W^n = 0] \\ &\quad + \sum_{v,w\in\mathcal{G}\backslash\{u\}} \mathbb{P}[K_{v,w}^n(\cdot) \neq 0 | Z^n(\tau_n) = 0]. \qquad (3.20)\end{aligned}$$

We will show that each term of the right hand side of Inequality (3.20) converges to zero. Firstly,

$$\mathbb{P}[W^n = 0 | Z^n(\tau_n) = n] = \left(\frac{\beta_n}{\alpha_n}\right)^n,$$

which is the probability that $Z^n(\cdot)$, if starting at size n, eventually goes extinct; this clearly converges to zero.

Secondly,

$$\begin{aligned}\mathbb{E}\left[\sup_t K_v^n(t) \middle| W^n = 0\right] &= \mathbb{E}\left[N_v^n\left(2p_n(u,v)\alpha(u')\int_0^\infty Z^n(s)ds\right)\middle| W^n = 0\right] \\ &= \mathbb{E}\Bigg[\mathbb{E}\Bigg[N_v^n\left(2p_n(u,v)\alpha(u')\int_0^\infty Z^n(s)ds\right) \\ &\qquad\qquad \Big| Z^n(\cdot)\Bigg]\Big| W^n = 0\Bigg] \\ &= \mathbb{E}\left[2p_n(u,v)\alpha(u')\int_0^\infty Z^n(s)ds \middle| W^n = 0\right] \\ &= 2p_n(u,v)\alpha(u')\int_0^\infty \mathbb{E}[Z^n(s)|W^n = 0]ds \\ &= 2p_n(u,v)\alpha(u')\int_0^\infty e^{-\lambda_n s}ds \\ &\to 0,\end{aligned}$$

because $p_n(u,v) \to 0$. Hence

$$\mathbb{P}\left[\sup_t K_v^n(t) \neq 0 \middle| W^n = 0\right] \to 0.$$

Lastly,

$$\begin{aligned}\mathbb{P}[K_{v,w}^n(\cdot) \neq 0 | Z^n(\tau_n) = 0] &= \mathbb{P}[K_{v,w}^n(\tau_n) \neq 0 | Z^n(\tau_n) = 0] \\ &\leq \frac{\mathbb{P}[K_{v,w}^n(\tau_n) \neq 0]}{\mathbb{P}[Z^n(\tau_n) = 0]}.\end{aligned}$$

But $\mathbb{P}[K_{v,w}^n(\tau_n) \neq 0]$ converges to zero by Lemma 3.6.9, while $\mathbb{P}[Z^n(\tau_n) = 0]$ converges to $\mathbb{P}[W^* = 0] > 0$ by (3.12). □

Proof of Part 2 of Proposition 3.6.15. Just as for Part 1,

$$\begin{aligned}\mathbb{P}[W^* = 0, \sigma_n < \infty] &= \mathbb{P}[W^* = 0, \sigma_n < \infty, W^n > 0] \\ &\quad + \mathbb{P}[W^* = 0, \sigma_n < \infty, W^n = 0] \\ &\leq \mathbb{P}[W^* = 0, W^n < 0] \\ &\quad + \mathbb{P}[\sigma_n < \infty, W^n = 0] \to 0\end{aligned}$$

as $n \to \infty$, by Lemmas 3.6.18 and 3.6.19. □

Corollary 3.6.20 (to Proposition 3.6.15)**.** *For any event H,*

$$\lim_{n\to\infty} \mathbb{P}[H, \sigma_n < \infty] = \mathbb{P}[H, W^* > 0].$$

Proof. Write

$$\mathbb{P}[H, \sigma_n < \infty] + \mathbb{P}[H, W^* > 0, \sigma_n = \infty] = \mathbb{P}[H, W^* > 0] + \mathbb{P}[H, W^* = 0, \sigma_n < \infty],$$

and take $n \to \infty$. □

Finally we are in a position to prove Theorem 3.6.1.

Proof of Theorem 3.6.1. For any $R \subset (\mathbb{N}_0 \cup \{\infty\})^{\mathcal{G}}$,

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathbb{P}[\mathcal{X}^n(\sigma_n) \in R, \sigma_n < \infty]}{\mathbb{P}[\sigma_n < \infty]} &= \lim_{n\to\infty} \frac{\mathbb{P}[\mathcal{X}^n(\sigma_n) \in R, W^* > 0]}{\mathbb{P}[W^* > 0]} \quad (3.21) \\ &= \mathbb{P}[(X_v)_{v\in\mathcal{G}} \in R], \quad (3.22)\end{aligned}$$

where (3.21) is due to Corollary 3.6.20 and (3.22) is due to Lemma 3.6.14. □

3.6.2 Corollaries 3.3.2 and 3.5.4

It will be clearer to write the site frequency spectrum as a sum of indicator functions:

$$|\{i \in \mathcal{S} : B_i^{n,\mu} = k\}| = \sum_{i\in\mathcal{S}} 1_{\{B_i^{n,\mu}=k\}}.$$

The expected site frequency spectrum, normalised, is

$$\begin{aligned}\mathbb{E}\left[|\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}1_{\{B_i^{n,\mu}=k\}}\right] &= |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}\mathbb{P}[B_i^{n,\mu}=k] \\ &= |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}_{\text{sel}}}\mathbb{P}[B_i^{n,\mu}=k] \\ &\quad +|\mathcal{S}|^{-1}\sum_{j\in J}\sum_{i\in\mathcal{S}(j)}\mathbb{P}[B_i^{n,\mu}=k].\end{aligned}$$

Theorem 3.5.3 says that $\mathbb{P}[B_i^{n,\mu}=k]\to\mathbb{P}[B(j)=k]$, while $|\mathcal{S}_{\text{sel}}|/|\mathcal{S}|\to 0$ and $|\mathcal{S}(j)|/|\mathcal{S}|\to q(j)$. Hence

$$\mathbb{E}\left[|\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}1_{\{B_i^{n,\mu}=k\}}\right]\to\sum_{j\in J}q(j)\mathbb{P}[B(j)=k].$$

The variance is

$$\begin{aligned}\text{Var}\left[|\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}1_{\{B_i^{n,\mu}=k\}}\right] &= |\mathcal{S}|^{-2}\sum_{i\in\mathcal{S}}\text{Var}[1_{\{B_i^{n,\mu}=k\}}] \\ &\quad +|\mathcal{S}|^{-2}\sum_{\substack{i,j\in\mathcal{S}\\ i\neq j}}\text{Cov}[1_{\{B_i^{n,\mu}=k\}},1_{\{B_j^{n,\mu}=k\}}] \\ &\leq |\mathcal{S}|^{-1}+\max_{\substack{i,j\in\mathcal{S}\\ i\neq j}}\text{Cov}[1_{\{B_i^{n,\mu}=k\}},1_{\{B_j^{n,\mu}=k\}}].\end{aligned}$$

Because $\mathcal{S}_{\text{sel}}$ and J are fixed sets and the random variables are exchangable over $\mathcal{S}(j)$, the maximum is taken over a fixed set. Theorem 3.5.3 says that the covariances converge to zero. □

3.6.3 Theorem 3.4.1

Because this result assumes no cell death,

$$(B_i^{n,\mu})_{n\in\mathbb{N}}$$

is a Markov process on the integers with transition probabilities

$$\begin{aligned}&\mathbb{P}[B_i^{n+1,\mu}=k|B_i^{n+1,\mu}=j] \\ &\quad=\begin{cases}\frac{j}{n}(\mu/6)^2, & k=j-1;\\ \frac{j}{n}2(\mu/6)(1-\mu/6)+\frac{n-j}{n}(1-\mu/2)^2, & k=j;\\ \frac{j}{n}(1-\mu/6)^2+\frac{n-j}{n}2(\mu/2)(1-\mu/2), & k=j+1;\\ \frac{n-j}{n}(\mu/2)^2, & k=j+2.\end{cases}\end{aligned} \tag{3.23}$$

The subscript i plays no part in this result, so for the proof let's drop i from the notation. The key idea of the proof will be to condition on the number of cells

when the first mutant (with respect to site i) arises. For this purpose, introduce

$$\xi^{\mu} = \min\{n \in \mathbb{N} : B^{n,\mu} > 0\}$$

for the total number of cells when the first mutant cell arises. For $k \in \{1, 2\}$, let

$$\Xi_k^{\mu} = \{B^{\xi^{\mu},\mu} = k\}$$

be the event that the first mutation gives rise to k mutant cells. Now, for $a \in (0, 1)$,

$$\begin{aligned}\mu^{-1}\mathbb{P}[B^{n,\mu} > an] \quad &= \quad \sum_{j=2}^{\infty}\sum_{k=1}^{2}\Big(\mathbb{P}[B^{n,\mu} > an \big| \xi^{\mu} = j, \Xi_k^{\mu}] \\ &\qquad \times \mu^{-1}\mathbb{P}[\xi^{\mu} = j, \Xi_k^{\mu}]\Big). \end{aligned} \tag{3.24}$$

Let $(\mu_n)_{n\in\mathbb{N}}$ be a positive sequence with

$$\lim_{n\to\infty} n\mu_n = \theta.$$

To see the limit of $\mu_n^{-1}\mathbb{P}[B^{n,\mu_n} > an]$, the following three lemmas will be applied to (3.24).

Lemma 3.6.21.

$$\lim_{n\to\infty} \mathbb{P}[B^{n,\mu_n} > an \big| \xi^{\mu_n} = j, \Xi_1^{\mu_n}] = (1-a)^{j-1}.$$

Lemma 3.6.22.

$$\lim_{n\to\infty} \mu_n^{-1}\mathbb{P}[\xi^{\mu_n} = j, \Xi_k^{\mu_n}] = \begin{cases} 1, & k = 1; \\ 0, & k = 2 \end{cases}.$$

Lemma 3.6.23.

$$\sup_{n\in\mathbb{N}} \mathbb{P}[B^{n,\mu_n} > an \big| \xi^{\mu_n} = j, \Xi_k^{\mu_n}] \le cj^{-2},$$

where $c > 0$ does not depend on j, k.

Proof of Theorem 3.4.1. Lemmas 3.6.21, 3.6.22, and 3.6.23, along with the Dominated Convergence Theorem, show that the limit of (3.24) is

$$\sum_{j=1}^{\infty}(1-a)^j = a^{-1} - 1.$$

□

It remains to prove Lemmas 3.6.21, 3.6.22, and 3.6.23, which is the subject of the remainder of this section. Let's start with Lemma 3.6.21. Its proof makes use of the construction in Section 3.6.1; notation is taken from there.

Lemma 3.6.24. *Suppose that* (a_n) *is a sequence with*

$$\lim_{n\to\infty} a_n = \infty$$

and

$$a_n \leq \sigma_n$$

almost surely. Then

$$\lim_{n\to\infty} e^{-\lambda_n a_n} \mathcal{X}^n(a_n) = W^* e_u,$$

almost surely.

Proof. Lemma 3.6.12 teaches us that $|\mathcal{X}^n(a_n)| - Z^n(a_n)$ is bounded, and therefore

$$\lim_{n\to\infty} \frac{|\mathcal{X}^n(a_n)|}{Z^n(a_n)} = 1.$$

Then, writing $\mathcal{X}^n_v(\cdot)$ for the vth entry of $\mathcal{X}^n(\cdot)$,

$$e^{-\lambda_n a_n} Z^n(a_n) \leq e^{-\lambda_n a_n} \mathcal{X}^n_u(a_n) \leq e^{-\lambda_n a_n} Z^n(a_n) \frac{|\mathcal{X}^n(a_n)|}{Z^n(a_n)}.$$

The upper and lower bounds both converge to W^* by (3.12), as required. As for $v \neq u$,

$$e^{-\lambda_n a_n} \mathcal{X}^n_v(a_n) \leq e^{-\lambda_n a_n} \left(|\mathcal{X}^n(a_n)| - Z^n(a_n) \right),$$

which converges to zero. □

Proof of Lemma 3.6.21. Let

$$\left(\mathcal{X}^{n,l}(\cdot)\right)_{n\in\mathbb{N}}$$

for $l = 1, .., j$, be independent copies of

$$\left(\mathcal{X}^n(\cdot)\right)_{n\in\mathbb{N}},$$

but with initial conditions

$$\mathcal{X}^{n,l}(0) = e_{u[l]}$$

for some genomes $u[1], .., u[j] \in \mathcal{G}$. Let

$$\sigma'_n = \min\left\{ t \geq 0 : \sum_{l=1}^{j} |\mathcal{X}^{n,l}(t)| = n \right\},$$

for $n \geq j$. Then σ'_n converges to infinity and

$$\sigma'_n \leq \sigma_{n,l} := \min\{t \geq 0 : |\mathcal{X}^{n,l}(t)| = n\}.$$

Therefore, using Lemma 3.6.24

$$\frac{\sum_{l=1}^{j} \mathcal{X}^{n,l}(\sigma'_n)}{\left|\sum_{l=1}^{j} \mathcal{X}^{n,l}(\sigma'_n)\right|} \to \frac{\sum_{l=1}^{j} W^{*,l} e_{u[l]}}{\sum_{l=1}^{j} W^{*,l}},$$

where the $W^{*,l}$ are independent copies of W^*, which is an Exp(1) random variable.

The statement of Lemma 3.6.21 asks that we take

$$u[l] = \begin{cases} u, & l = 1, .., j-1; \\ v, & l = j; \end{cases}$$

where v is some genome which is mutated at site i. Then

$$\left(n^{-1}B^{n,\mu_n} | \xi^{\mu_n} = j, M_1^{\mu_n}\right) \stackrel{d}{=} \frac{\sum_{l=1}^{j} \mathcal{X}_v^{n,l}(\sigma_n')}{\left|\sum_{l=1}^{j} \mathcal{X}^{n,l}(\sigma_n')\right|} \to \frac{W^{*,j}}{\sum_{l=1}^{j} W^{*,l}}.$$

But the limit is just a beta random variable, giving the result. □

Proof of Lemma 3.6.22. The probability that the first $j-2$ cell divisions give no site i mutations multiplied by the probability that the $(j-1)$th cell division gives exactly one mutant daughter is

$$\mathbb{P}[\xi^{\mu_n} = j, \Xi_1^{\mu_n}] = (1 - \mu_n/2)^{2j-3}\mu_n.$$

Similarly

$$\mathbb{P}[\xi^{\mu_n} = j, \Xi_2^{\mu_n}] = (1 - \mu_n/2)^{2j-4}\mu_n^2/4.$$

The result is immediate. □

Proof of Lemma 3.6.23. Let $r \in \{2, .., n\}$. Note that $n\mu_n$ is bounded above and that $B^{r,\mu_n} \leq r$. Directly calculating from the transition probabilities (3.23),

$$\mathbb{E}[(B^{r+1,\mu_n})^2 | B^{r,\mu_n}] \leq (r+1)^2 r^{-2} (B^{r,\mu_n})^2 + b,$$

where $b > 0$ is some constant independent of n, r. Taking expectations and rearranging,

$$\mathbb{E}[(B^{r+1,\mu_n}/(r+1))^2] - \mathbb{E}[(B^{r,\mu_n}/r)^2] \leq r^{-2}b.$$

This leads to

$$\mathbb{E}[(B^{r,\mu_n}/r)^2] \leq \mathbb{E}[(B^{j,\mu_n}/j)^2] + b\sum_{k=j}^{r-1} k^{-2}. \tag{3.25}$$

The lemma which we are proving asks to condition the Markov chain $(B^{r,\mu_n})_{r \geq j}$ on $\{\xi^{\mu_n} = j, \Xi_k^{\mu_n}\}$, which is just conditioning on the initial state $B^{j,\mu_n} = k$. To keep notation brief, let's condition on this initial state without explicitly writing conditional expectations. Then (3.25) becomes

$$\begin{aligned} \mathbb{E}[(B^{r,\mu_n}/r)^2] &\leq k^2 j^{-2} + b\sum_{k=j}^{r-1} k^{-2} \\ &\leq b' j^{-1}, \end{aligned}$$

where b' is another constant. (In the following, b'' and b''' will also be constants.)

A similar calculation for third moments gives

$$\begin{aligned}\mathbb{E}[(B^{r+1,\mu_n}/(r+1))^3] - \mathbb{E}[(B^{r,\mu_n}/r)^3] &\leq b''r^{-4}\mathbb{E}[(B^{r,\mu_n}/r)^2] \\ &\leq b'''r^{-2}j^{-1},\end{aligned}$$

from which it follows that

$$\begin{aligned}\mathbb{E}[(B^{n,\mu_n}/n)^3] &= \mathbb{E}[(B^{j,\mu_n}/j)^3] + b'''j^{-1}\sum_{r=j}^{n-1} r^{-2} \\ &\leq j^{-3} + b'''j^{-2} \\ &\leq cj^{-2}.\end{aligned}$$

Apply Markov's inequality to conclude. □

3.6.4 Theorem 3.4.4

Evolutionary tree

To understand the dependency structure of large-frequency mutations, one first needs to understand, to an extent, the evolutionary tree. Recall the notation for the tree which was introduced in Section 3.4:

$$\mathcal{T} = \cup_{l=0}^{\infty}\{0,1\}^l,$$

and its partial ordering '$\prec$' denoting ancestral relationships. Now further notation is introduced.

The lifetimes of the cells are given by i.i.d. Exp(α) random variables

$$(A_x)_{x\in\mathcal{T}},$$

and the cells alive at time t are

$$\mathcal{T}_t = \{x \in \mathcal{T} : \sum_{y\prec x} A_y \leq t < \sum_{y\preceq x} A_y\}.$$

In this notation,

$$\sigma_n = \min\{t \geq 0 : |\mathcal{T}_t| = n\}.$$

Write

$$\mathcal{D}_x = \{y \in \mathcal{T} : x \preceq y\}$$

for the descendants of cell x, and then

$$\mathcal{D}_{x,t} = \mathcal{D}_x \cap \mathcal{T}_t$$

for the cells alive at time t which are descendants of x (for convenience we say that x is a descendant of x). Write

$$P_{x,t} = \frac{|\mathcal{D}_{x,t}|}{|\mathcal{D}_{\emptyset,t}|} = \frac{|\mathcal{D}_{x,t}|}{|\mathcal{T}_t|}$$

for the proportion of cells alive at time t which are descendants of x.

The following result states the long term proportion of descendants of each cell. The result is surely not new, but we do not know of a reference.

Lemma 3.6.25. *For each $x \in \mathcal{T}$,*

$$\lim_{t\to\infty} P_{x,t} = P_x := \prod_{\emptyset \prec y \preceq x} U_y$$

almost surely, where

1. *the U_y are uniform random variables on $[0,1]$;*
2. *for any $y \in \mathcal{T}$, $U_{y0} + U_{y1} = 1$;*
3. *$(U_{y0})_{y\in\mathcal{T}}$ is an independent family.*

Proof. Let $x \in \mathcal{T}$. Then

$$\begin{aligned}\mathcal{D}_{x,\sum_{y\prec x} A_y + t} &= \left\{ y \in \mathcal{T} : x \preceq y, \sum_{z\prec y} A_z \le \sum_{z\prec x} A_z + t < \sum_{z\preceq y} A_z \right\} \\ &= \left\{ y \in \mathcal{T} : x \preceq y, \sum_{x\preceq z\prec y} A_z \le t < \sum_{x\preceq z\preceq y} A_z \right\}.\end{aligned}$$

Hence

$$\left(|\mathcal{D}_{x,\sum_{y\prec x} A_y + t}| \right)_{t\ge 0}$$

is measurable with respect to the sigma-algebra generated by $(A_y)_{y\in\mathcal{D}_x}$, and has the same distribution as

$$(|\mathcal{D}_{\emptyset,t}|)_{t\ge 0} = (|\mathcal{T}_t|)_{t\ge 0}\,.$$

It follows that

$$\lim_{t\to\infty} e^{-\alpha t} |\mathcal{D}_{x,\sum_{y\prec x} A_y + t}| = \lim_{t\to\infty} e^{-\alpha t + \sum_{y\prec x} A_y} |\mathcal{D}_{x,t}| =: W_x$$

almost surely, where $W_x \sim$Exp(1). Moreover, if $x, y \in \mathcal{T}$ are such that $\mathcal{D}_x \cap \mathcal{D}_y = \emptyset$, then W_x and W_y are independent. In particular, W_{x0} and W_{x1} are independent. Now,

$$\lim_{t\to\infty} \frac{|\mathcal{D}_{x0,t}|}{|\mathcal{D}_{x,t}|} = \lim_{t\to\infty} \frac{|\mathcal{D}_{x0,t}|}{1 + |\mathcal{D}_{x0,t}| + |\mathcal{D}_{x1,t}|} = \frac{W_{x0}}{W_{x0} + W_{x1}} =: U_{x0}$$

almost surely, and $U_{x0} + U_{x1} = 1$. A standard calculation shows that U_{x0} is uniformly distributed on $[0,1]$: for $u \in (0,1)$,

$$\mathbb{P}[U_{x0} < u] = \int_0^\infty \int_{z(1-u)/u}^\infty e^{-y} e^{-z} dy dz = u.$$

It remains to show independence of the U_{x0}. Another standard calculation shows that

$$U_{x0} = \frac{W_{x0}}{W_{x0} + W_{x1}}$$

is independent of

$$W_{x0} + W_{x1} :$$

for $(u, v) \in (0, 1) \times (0, \infty)$,

$$\begin{aligned} \mathbb{P}[U_{x0} < u, W_{x0} + W_{x1} < v] &= \int_0^{uv} \int_{z(1-u)/u}^{v-z} e^{-y} e^{-z} dy dz \\ &= u(1 - (1 + v)e^{-v}) \\ &= \mathbb{P}[U_{x0} < u]\mathbb{P}[W_{x0} + W_{x1} < v]. \end{aligned}$$

Now fix $l \in \mathbb{N}$. Because U_{x0} and $W_{x0} + W_{x1}$ are measurable with respect to the sigma-algebra generated by $(A_y)_{y \in \mathcal{D}_x \setminus \{x\}}$, we have that

$$\left[(U_{x0})_{|x|=l}, (W_{x0} + W_{x1})_{|x|=l}, (A_x)_{|x| \le l}\right] \tag{3.26}$$

forms an independent collection of random variables.

Finally we complete the proof by induction. Suppose that $(U_{x0})_{x \in \mathcal{T} : |x| < l}$ is an independent family. Observing that for any $x \in \mathcal{T}$,

$$W_x = e^{-A_x}(W_{x0} + W_{x1}),$$

we have that $(U_{x0})_{x \in \mathcal{T} : |x| < l}$ is measurable with respect to the sigma-algebra generated by

$$[(W_{x0} + W_{x1})_{|x|=l}, (A_x)_{|x| \le l}].$$

Then, thanks to the independence of (3.26), $(U_{x0})_{x \in \mathcal{T} : |x| \le l}$ forms an independent family. □

Infinite-sites approximation

Having understood a little more of the evolutionary tree, the proof of Theorem 3.4.4 can really begin. The key idea is to consider an 'infinite-sites' version of the process. By infinite-sites we are not talking about taking the number of sites to infinity; we are referring to the infinite-sites assumption, where parallel and backward mutations are neglected. Under the assumption, any mutation arising in a cell will appear in all of its descendants and in no other cells. The proof of Theorem 3.4.4 can be summarised as: mutation frequencies are established under the infinite-sites assumption, which is shown provide a good approximation.

Let (μ_n) be a sequence of mutation rates with $n\mu_n \to \theta$, and $(\mathcal{S}_n)$ be a sequence of sets of sites with $|\mathcal{S}_n|\mu_n \to \eta$.

Fix $n \in \mathbb{N}$. Write $V^n(x) = (V_i^n(x))_{i \in \mathcal{S}_n}$ for the genetic state of cell $x \in \mathcal{T}$. So $(V_i^n(x))_{x \in \mathcal{T}}$ is a Markov process which is indexed by the tree $\mathcal{T}$ and takes values in $\mathcal{N}^{\mathcal{S}_n}$. And for $i \in \mathcal{S}_n$, the $(V_i^n(x))_{x \in \mathcal{T}}$ are independent $\mathcal{N}$-valued Markov processes.

Enumerate the elements of $\mathcal{T}$,

$$\mathcal{T} = (x_k)_{k \in \mathbb{N}},$$

in such a way that

$$x_j \prec x_k \implies j < k. \tag{3.27}$$

Let's give an example of such an enumeration: map $(x(r))_{r=1}^l \in \{0,1\}^l \subset \mathcal{T}$ to $2^l + \sum_{r=1}^l 2^{r-1}x(r)$.

Write

$$\phi_i^n = \min\{x \in \mathcal{T} : V_i^n(x) \neq u_i\}$$

for the first cell which sees a mutation at site i (where the minimum is with respect to the enumeration of $\mathcal{T}$). Write

$$M_x^n = |\{i \in \mathcal{S}_n : \phi_i^n = x\}|$$

for the number of sites which see their first mutation at cell x.

Lemma 3.6.26.

$$\lim_{n\to\infty} (M_x^n)_{x\in\mathcal{T}\setminus\{\emptyset\}} = (M_x)_{x\in\mathcal{T}\setminus\{\emptyset\}}$$

in distribution, where the M_x are i.i.d. Poisson($\eta/2$) random variables.

Proof. The initial cell is $x_1 = \emptyset$. The number of sites which mutate in cell x_2, $M_{x_2}^n$, is binomially distributed with parameters $\mathcal{S}_n$ and $(1-\mu_n/2)\mu_n/2$. This converges to a Poisson($\eta/2$) random variable. Now, for induction, suppose that

$$\lim_{n\to\infty} \left(M_{x_j}^n\right)_{j=2}^k = \left(M_{x_j}\right)_{j=2}^k$$

in distribution, where the M_x are i.i.d. Poisson($\eta/2$) random variables. Then

$$\begin{aligned}\mathbb{P}\left[\left(M_{x_j}^n\right)_{j=2}^{k+1} = (m_j)_{j=2}^{k+1}\right] &= \mathbb{P}\left[M_{x_{k+1}}^n = m_{k+1} \,\middle|\, \left(M_{x_j}^n\right)_{j=2}^k = (m_j)_{j=2}^k\right] \\ &\quad \times \mathbb{P}\left[\left(M_{x_j}^n\right)_{j=2}^k = (m_j)_{j=2}^k\right]. \end{aligned} \tag{3.28}$$

Due to the property (3.27) of the enumeration, $M_{x_{k+1}}^n$ conditioned on the event $\left(M_{x_j}^n\right)_{j=2}^k = (m_j)_{j=2}^k$ is just a binomial random variable with parameters $|\mathcal{S}_n| - \sum_{j=2}^k m_j$ and $(1-\mu_n/2)\mu_n/2$. Therefore (3.28) converges as required. □

Proposition 3.6.27. *For $a \in (0,1)$,*

$$\lim_{n\to\infty} |\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}| = \sum_{x\in\mathcal{T}\setminus\{\emptyset\}} M_x 1_{\{P_x > a\}}$$

in distribution.

Proof. From Lemmas 3.6.25 and 3.6.26, and the fact that the (M_x^n) are independent of the (P_{x,σ_n}):

$$\lim_{n\to\infty} (M_x^n, P_{x,\sigma_n})_{x\in\mathcal{T}\setminus\{\emptyset\}} = (M_x, P_x)_{x\in\mathcal{T}\setminus\{\emptyset\}}$$

in distribution (where the M_x are independent of the P_x). Then, by the continuous mapping theorem,

$$|\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}| = \sum_{x \in \mathcal{T}\setminus\{\emptyset\}} M_x^n 1_{\{P_{x,\sigma_n} > a\}}$$

converges as required. □

Write

$$\mathcal{B}_i^n = \{x \in \mathcal{T} : V_i^n(x) \neq u_i\}$$

for the cells which are mutated at site i, and

$$\hat{\mathcal{B}}_i^n = \cup_{x \in \mathcal{B}_i^n} \mathcal{D}_x$$

for their descendants (recall that $\mathcal{D}_x$ are the descendants of x). Note that

$$\mathcal{B}_i^n \subset \hat{\mathcal{B}}_i^n,$$

and then

$$|\mathcal{B}_i^n \cap \mathcal{T}_{\sigma_n}| \leq |\hat{\mathcal{B}}_i^n \cap \mathcal{T}_{\sigma_n}| =: \hat{B}_i^n,$$

where $\hat{B}_i^n$ is the number of cells alive at time σ_n which have descended from a mutant. Let's connect with the original notation:

$$B_i^{n,\mu_n} = |\mathcal{B}_i^n \cap \mathcal{T}_{\sigma_n}|$$

is just the number of cells alive at time σ_n which are mutated at site i, and

$$B_i^{n,\mu_n} \leq \hat{B}_i^n.$$

It follows that

$$\{i \in \mathcal{S}_n : B_i^{n,\mu_n} > an\} \subset \{i \in \mathcal{S}_n : \hat{B}_i^n > an\}. \tag{3.29}$$

We will come back to (3.29) at the end of Theorem 3.4.4's proof.

Now let's look at the descendants of the first mutant cell, which are clearly a subcollection of the descendants of all mutant cells:

$$\mathcal{D}_{\phi_i^n} \subset \hat{\mathcal{B}}_i^n.$$

And then

$$\begin{aligned} nP_{\phi_i^n,\sigma_n} &= |\mathcal{D}_{\phi_i^n} \cap \mathcal{T}_{\sigma_n}| \\ &\leq |\hat{\mathcal{B}}_i^n \cap \mathcal{T}_{\sigma_n}| \\ &= \hat{B}_i^n. \end{aligned}$$

Therefore

$$\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\} \subset \{i \in \mathcal{S}_n : \hat{B}_i^n > an\}. \tag{3.30}$$

The following two lemmas show that the expected sizes of the sets in (3.30) converge to the same limit.

Lemma 3.6.28.

$$\lim_{n\to\infty} \mathbb{E}|\{i \in \mathcal{S}_n : \hat{B}_i^n > an\}| = \eta(a^{-1} - 1).$$

Proof. Let $n \in \mathbb{N}, i \in \mathcal{S}_n$. Set

$$\hat{V}_i^n(x) = u_i$$

for $x \in \mathcal{T}\backslash\hat{\mathcal{B}}_i^n$, and

$$\hat{V}_i^n(xr) = \hat{V}_i^n(x)$$

for $x \in \hat{\mathcal{B}}_i^n, r \in \{0, 1\}$. Then

$$(\hat{V}_i^n(x))_{x\in\mathcal{T}}$$

is a Markov process indexed by $\mathcal{T}$ which takes values in $\mathcal{N}$. Initially

$$\hat{V}_i^n(\emptyset) = u_i,$$

and the process has transition probabilities

$$\begin{aligned}
&\mathbb{P}[\hat{V}_i^n(xr) = \psi | \hat{V}_i^n(x) = \chi] \\
&= \begin{cases} (1 - \mu_n/2), & \chi = \psi = u_i; \\ \mu_n/6, & \chi = u_i, \psi \neq u_i; \\ 1, & \chi \neq u_i, \chi = \psi; \\ 0, & \chi \neq u_i, \chi \neq \psi. \end{cases}
\end{aligned} \tag{3.31}$$

Also, the processes $(\hat{V}_i^n(x))_{x\in\mathcal{T}}$ are independent over $i \in \mathcal{S}_n$.

Note that

$$\hat{B}_i^n = |\{x \in \mathcal{T}_{\sigma_n} : \hat{V}_i^n(x) \neq u_i\}|$$

appears very similar to the quantity

$$B_i^{n,\mu_n} = |\{x \in \mathcal{T}_{\sigma_n} : V_i^n(x) \neq u_i\}|.$$

In fact, we can consider the $\hat{V}_i^n(x)$ as an alternative model for genetic information which is a special case of the general heterogeneous mutation rate setting laid out in Section 3.5.3. To connect the notation of Section 3.5.3 and (3.31),

$$\mu_{i,n}^{\psi,\chi}/2 = \mathbb{P}[\hat{V}_i^n(xr) = \psi | \hat{V}_i^n(x) = \chi].$$

In this model, $\hat{B}_i^n$ is just the number of site i mutants when the total population size reaches n. Crucially,

$$\lim_{n\to\infty} n\mu_{n,i}^{\chi,\psi} = \theta_i^{\chi,\psi},$$

where

$$\sum_{\psi\in\mathcal{N}\backslash\{u_i\}} \theta_i^{\chi,\psi} = \theta.$$

Therefore the proof of this Lemma boils down to proving Theorem 3.4.1 in a slightly different, heterogeneous mutation rate setting. The proof works almost identically in both settings, and we do not reproduce it. □

Lemma 3.6.29.

$$\lim_{n\to\infty} \mathbb{E}|\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}| = \eta(a^{-1} - 1).$$

Proof. By (3.30),

$$|\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}| \leq |\{i \in \mathcal{S}_n : \hat{B}_i^n > an\}|.$$

Therefore, by Lemma 3.6.28,

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{E}|\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}| &\leq \lim_{n\to\infty} |\{i \in \mathcal{S}_n : \hat{B}_i^n > an\}| \\ &= \eta(a^{-1} - 1). \end{aligned}$$

Meanwhile, by Fatou's lemma and Proposition 3.6.27,

$$\begin{aligned} \liminf_{n\to\infty} \mathbb{E}|\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}| &\geq \mathbb{E} \sum_{x\in\mathcal{T}\setminus\{\emptyset\}} M_x 1_{\{P_x > a\}} \\ &= \eta(a^{-1} - 1). \end{aligned}$$

□

Finally the threads can be tied together to complete the proof of Theorem 3.4.4.

Proof of of Theorem 3.4.4. Write

$$\zeta^n = |\{i \in \mathcal{S}_n : B_i^{n,\mu_n} > an\}|,$$

$$\hat{\zeta}^n = |\{i \in \mathcal{S}_n : \hat{B}_i^n > an\}|,$$

and

$$\zeta_P^n = |\{i \in \mathcal{S}_n : P_{\phi_i^n,\sigma_n} > a\}|.$$

Then

$$\begin{aligned} \mathbb{E}|\zeta^n - \zeta_P^n| &\leq \mathbb{E}|\hat{\zeta}^n - \zeta^n| + \mathbb{E}|\hat{\zeta}^n - \zeta_P^n| \\ &= 2\mathbb{E}\hat{\zeta}^n - \mathbb{E}\zeta^n - \mathbb{E}\zeta_P^n \\ &\to 0, \end{aligned} \tag{3.32}$$

where the equality is due to (3.29) and (3.30), and the convergence to zero is due to Lemmas 3.6.28 and 3.6.29 and Theorem 3.4.1. Proposition 3.6.27 gives the limiting distribution of ζ_P^n. By (3.32), ζ^n and ζ_P^n share the same limiting distribution. □

3.7 Support for conjectures

A heuristic derivation of Conjectures 3.5.5 and 3.5.6 is given.

First we argue that, in the conjectures' limit, selection is unimportant. Write

$$\mathcal{G}_{\text{sel}} = \{v \in \mathcal{G} : \exists i \in \mathcal{S}_{\text{sel}}, v_i \neq u_i\}$$

for the set of genomes which are mutated at a selective site. Write

$$Q^{\mu}_{\text{sel}}(t) = \frac{\sum_{v \in \mathcal{G}_{\text{sel}}} X^{\mu}_v(t)}{\sum_{v \in \mathcal{G}} X^{\mu}_v(t)}$$

for the proportion of cells at time $t \geq 0$ whose genomes are mutated at a selective site. Then, according to Theorem 3.6.1,

$$(Q^{\mu}_{\text{sel}}(\sigma^{\mu}_n) | \sigma^{\mu}_n < \infty) \to 0$$

in distribution. Therefore we neglect selection.

Next let's discuss cell death. Some cells have a long-term surviving lineage of descendants, while other cells eventually have no surviving descendants. Name these cells immortal and mortal respectively. In a supercritical birth-death branching process, it is well-known (eg. [10]) that the immortal cells grow as a Yule process and the mortal cells grow as a subcritical branching process. An immortal cell divides to produce two immortal cells at rate $\alpha - \beta$, or it divides to produce one immortal and one mortal cell at rate 2β. A mortal cell divides at rate β to produce two mortal cells, or it dies at rate α. Assume that the initial cell is immortal.

The tree notation of Section 3.4, $\mathcal{T} = \cup_{l=0}^{\infty} \{0,1\}^l$ and its partial ordering $\prec$, will be used to represent the immortal cells. Let $(A_x)_{x \in \mathcal{T}}$ be i.i.d. $\text{Exp}(\alpha - \beta)$ random variables, which represent the times for immortal cells to divide to produce two immortal cells. The immortal cells at time $t \geq 0$ are

$$\mathcal{T}_t = \{x \in \mathcal{T} : \sum_{y \prec x} A_y \leq t < \sum_{y \preceq x} A_y\}.$$

The immortal descendants of $x \in \mathcal{T}$ are

$$\mathcal{D}_x = \{y \in \mathcal{T} : x \preceq y\}.$$

The number of immortal descendants of cell x at time t is

$$D^I_{x,t} = |\mathcal{D}_x \cap \mathcal{T}_t|.$$

Let $((R_x(t))_{t \geq 0})_{x \in \mathcal{T}}$ be i.i.d. Poisson processes with rate 2β. Write $R_{x,i} = \min\{t \geq 0 : R_x(t) = i\}$ for $i = 1, .., R_x(A_x)$. Then the seeding times of mortal cells are

$$S_{x,i} = \sum_{y \prec x} A_y + R_{x,i}.$$

Each seeding event initiates a subpopulation of mortal cells; let $(Y_{x,i}(t))_{t\geq 0}$ be i.i.d. birth-death branching processes with birth and death rates β and α. Then the number of mortal descendants of x at time t is

$$D^M_{x,t} = \sum_{y\in\mathcal{D}_x} \sum_{i=1}^{R_y(A_y)} 1_{\{t-S_{y,i}\geq 0\}} Y_{y,i}\left(t - S_{y,i}\right).$$

The number of descendants of x at time t is

$$D_{x,t} = D^I_{x,t} + D^M_{x,t}.$$

The next result shows the long-term proportion of a cell's descendants which are immortal. The result is a basic consequence of classic branching process theory [4], and was mentioned in its specific form by [10].

Lemma 3.7.1. *The limit*

$$\lim_{t\to\infty} \frac{D^I_{x,t}}{D_{x,t}} \in (0,\infty)$$

exists almost surely, and is deterministic.

We use Lemma 3.7.1 to see the number of descendants of a cell as a proportion of the total population.

Lemma 3.7.2. *For $x \in \mathcal{T}$,*

$$\lim_{t\to\infty} \frac{D_{x,t}}{D_{\emptyset,t}} = P_x$$

almost surely, where the P_x are as in Lemma 3.6.25.

Proof. By Lemma 3.6.25 and Lemma 3.7.1,

$$\frac{D_{x,t}}{D_{\emptyset,t}} = \frac{D_{x,t}}{D^I_{x,t}} \frac{D^I_{x,t}}{D^I_{\emptyset,t}} \frac{D^I_{\emptyset,t}}{D_{\emptyset,t}}$$

converges to the required limit. □

Let's look at mutations. In the proof of Theorem 3.4.4 it was shown that the number of new mutations to arise at a cell's birth is approximately Poisson. Here, with heterogeneous mutation rates, the number of new mutations to arise at a cell's birth is approximately Poisson with mean

$$\eta/2 := \sum_{j\in J} \sum_{\psi\in\mathcal{N}\setminus\{u(j)\}} \eta^{u(j),\psi}(j)/2.$$

Each $x \in \mathcal{T}$ witnesses $1 + R_x(A_x)$ cell divisions. So the number of new mutations to arise at x is

$$\sum_{i=0}^{R_x(A_x)} M_{x,i}, \tag{3.33}$$

where the $M_{x,i}$ are i.i.d. Poisson random variables with mean $\eta/2$. In the proof of Theorem 3.4.4 it was also shown that a mutation which arises at $x \in \mathcal{T}$ will have approximate frequency P_x. Here, thanks to Lemma 3.7.2, the situation appears identical. But what happens to mutations arising in mortal cells? Any subpopulation of cells which descended from a mortal cell must eventually die out. Hence mutations arising in mortal cells are negligible when compared to the total population size (which tends to infinity). This concludes the heuristic argument for Conjecture 3.5.6.

For $\beta > 0$, the random variable A_x appears in both M_x and P_x, and hence M_x and P_x are not independent. Without independence, it is not straightforward to take the expectation of Conjecture 3.5.6's limit, to recover Conjecture 3.5.5. But a derivation for Conjecture 3.5.5 readily comes from the immortal-mortal decomposition.

Recall that Conjecture 3.5.5 generalises Corollary 3.4.3 to $\beta \geq 0$. Recall that the proof of Corollary 3.4.3 involves counting the expected number of mutations to arise when there are $k \in \mathbb{N}$ cells, and seeing the long-term proportion of each of the k cells' descendants. For $\beta > 0$, the situtation is only a little more complex. When there are k immortal cells, the expected number of immortal cell divisions which seed a mortal cell is $2\beta/(\alpha - \beta)$. Therefore, when there are k immortal cells, the expected number of sites to mutate in an immortal cell is

$$\eta + \frac{\eta}{2}\frac{2\beta}{\alpha - \beta} = \frac{\eta\alpha}{\alpha - \beta}. \tag{3.34}$$

(The first term of (3.34), η, corresponds to the division which took $k-1$ immortal cells to k immortal cells.) Recall that, according to Lemma 3.7.2, the long-term proportion of descendants of a particular immortal cell is indifferent to the parameter β. So the only impact of β is the factor $\alpha/(\alpha-\beta)$ in (3.34). Conjecture 3.5.5 simply incorporates this factor into Corollary 3.4.3.

Note that the tail distribution of Corollary 3.5.4's limit coincides with Conjecture 3.5.5 (recall the Luria-Delbrück distribution's power-law tail, seen in Remark 2.3.5). Note further that [39, 7, 10] derive the same result for their neutral infinite-sites models.

3.8 Connecting to data

Now we dip our toes into the arena of cancer genetic data. Our intention here is not novel nor in-depth statistical analysis. Instead we try to keep matters as simple as possible. We wish, with a single example, to give the reader a flavour of data's appearance and its relationship to the model.

3.8.1 Diploid perspective

Before presenting data, an additional ingredient needs to be considered: ploidy. Normal human cells are diploid. That is, chromosomes come in pairs. Therefore a particular mutation may be present zero, one, or two times in a single cell. It should be said that the story is far more complex in tumours, with chromosomal

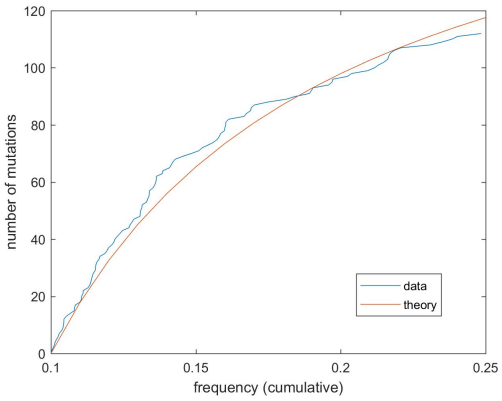


Figure 3.3: The number of mutations (of the lung adenocarcinoma) whose frequency is in the interval $(0.1, x)$, for $x \in (0.1, 0.25)$.

instability and aneuploidy coming into play. Even so, many tumour samples display an average ploidy not so far from two (for example see Figure (1a) of [40]). We imagine an idealised diploid world.

To illustrate the diploid structure, label the genetic sites as

$$\mathcal{S} = \{1, 2\} \times \{1, .., L\},$$

for some $L \in \mathbb{N}$. The first coordinate of a site $(i, j) \in \mathcal{S}$ states on which chromosome of a pair the site lies, and the second coordinate refers to the site's position on the chromosome. Mutations at sites $(1, j)$ and $(2, j)$ are typically not distinguished in data. In the original model set up, mutations were defined as differences to the initial cell's genome. Let's slightly improve that definition. Now a genome $v \in \mathcal{G}$ is said to be mutated at site $(i, j) \in \mathcal{S}$ if $v_{i,j} \neq r_j$, where $(r_j)_{j=1}^{L}$ is some reference. Then data is simplistically stated in the model's language as

$$F_j = \frac{1}{2n} \sum_{i=1}^{2} B_{i,j}^{n,\mu} \tag{3.35}$$

for $j = 1, .., L$. That is, the total number of mutations at position j divided by the total number of chromosomes which contain position j.

3.8.2 Example: lung adenocarcinoma

The mutation frequency data of a lung adenocarcinoma was made available in [17] (499017, Table S2). The data is plotted in Figures 1.1 and 3.3. It must be noted that a multitude of different data shapes exist, some looking nothing like this one. Our aim is only to give a flavour, hence just a single cancer's data is presented.

Our method to estimate mutation rates is, to a large extent, inspired by [39, 7]. Their attention is restricted to a subset of mutations. They ignore mutations at frequency less than 0.1, saying that their detection is too unreliable. They ignore

mutations above frequency 0.25, in order to neglect mutations present in the initial cell (which are few compared to the genome size). We do the same.

Write

$$\mathcal{M}(a,b) = |\{j \in \{1,..,L\} : F_j \in (a,b)\}|$$

for the number of mutations with frequency in $(a,b) \subset (0.1, 0.25)$. Then, adapting (3.6) to (3.35), the expected number of mutations with frequency in (a,b) is

$$\mathbb{E}\mathcal{M}(a,b) \approx \frac{1}{2}\mu|\mathcal{S}|(a^{-1} - b^{-1}). \tag{3.36}$$

Under different models, [39, 7] derive the same approximation (3.36). They estimate μ by applying a linear regression to (3.36). We simplify matters even further. Our estimator for μ is

$$\hat{\mu} = \frac{\mathcal{M}(0.1, 0.25)}{3|\mathcal{S}|}, \tag{3.37}$$

which (3.36) says is asymptotically unbiased. On the other hand, the variance of $\hat{\mu}$ is not insignificant. Similarly to Remark 3.4.6,

$$\mathrm{Var}[\hat{\mu}/\mu] \gtrapprox \frac{1}{3|\mathcal{S}|\mu},$$

which is apparently not so far from 1. Now let's calulate $\hat{\mu}$ for the example data. The data shows mutations on the exome, which has rough size $|\mathcal{S}| = 3 \times 10^8$ [7]. And the number of mutations in the specified frequency range is $\mathcal{M}(0.1, 0.25) = 112$. This gives

$$\hat{\mu} = 1.2 \times 10^{-7}.$$

Next let's consider mutation rate heterogeneity. Write μ_χ for the rate that nucleotide $\chi \in \mathcal{N}$ mutates at cell division. Partition the genetic sites:

$$\mathcal{S} = \mathcal{S}_A \cup \mathcal{S}_C \cup \mathcal{S}_G \cup \mathcal{S}_T,$$

where

$$\mathcal{S}_\chi = \{i \in \mathcal{S} : u_i = \chi\}$$

is the set of sites which are represented by nucleotide χ in the initial cell. Just as before,

$$\hat{\mu}_\chi = \frac{\mathcal{M}_\chi(0.1, 0.25)}{3|\mathcal{S}_\chi|}$$

is an unbiased estimator for μ_χ. The data gives

$$(\hat{\mu}_A, \hat{\mu}_C, \hat{\mu}_G, \hat{\mu}_T) = (0.1, 2.6, 3.0, 0.3) \times 10^{-7}.$$

What if cell death is included in the model? In this case the estimation story needs to be changed: (3.37) estimates

$$\frac{\alpha}{\alpha - \beta}\mu. \tag{3.38}$$

Interestingly, [7] turned (3.38) on its head. By taking an estimate for μ from the literature and inverting (3.38), they obtained estimates for β/α.

That such simple mathematical illustrations can illuminate otherwise obscure evolutionary processes is, we think, exciting. But needless to say, one should question the model's assumptions. On this note, numerous paths for future research are suggested. Let's mention one here which is especially important to the biology. What happens away from the world of branching processes, where cells do not divide and die independently?

One should also question the data itself. As discussed in [39], the data is affected by:

1. Spatial sampling - Only a fraction of the cells are taken from the tumour, and these are not independently selected. Samples are taken from distinct spatial regions in the tumour.

2. Impurity - The cancer cells are mixed together with normal cells from the surrounding tissue.

3. Noise in DNA reading - From extracting to reading DNA, there is a complex and technologically involved process with several stages. Throughout the process, noise is introduced.

We do not mean to bad-mouth data however. Quite the opposite. Mutation frequency data and other forms of cancer genetic data are becoming increasingly abundant, diverse, and precise. Biologists, statisticians, and mathematicians are coming together to interpret this growing library, developing ever deeper biological insights. It is clear that mathematical models will continue to play a role here; and their role is not just restricted to suggesting statistical inference methods. Models, in their idealised simplicity, can illustrate the interplay between key features of the biological processes. Finally we must thank biology for inspiring mathematics which, we believe, is interesting in its own right.

Bibliography

[1] D. Aldous. Stopping times and tightness. *Annals of Probability*, 6:335–340, 1978.

[2] W. P. Angerer. An explicit representation of the Luria-Delbrück distribution. *Journal of Mathematical Biology*, 42(2):145–174, 2001.

[3] T. Antal and P. L. Krapivsky. Exact solution of a two-type branching process: Models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, P08018, 2011.

[4] K. B. Arthreya and P. Ney. *Branching Processes*. Dover Publications, 2004.

[5] P. Billingsley. *Convergence of Probability Measures*. New York: Wiley, 1968.

[6] I. Bozic et al. Evolutionary dynamics of cancer in response to targeted combination therapy. *Elife*, (2):e00747, 2013.

[7] I. Bozic, J. M. Gerold, and M. A. Nowak. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLOS Computational Biology*, 12(2):e1004731, 2016.

[8] L. A. Diaz Jr et al. The molecular evolution of acquired resistance to targeted egfr blockade in colorectal cancers. *Nature*, 486(7404):537, 2012.

[9] D. Dingli et al. The emergence of tumor metastases. *Cancer Biology and Therapy*, 6(3):383–390, 2007.

[10] R. Durrett. Population genetics of neutral mutations in exponentially growing cancer cell populations. *The Annals of Applied Probability*, 23(1):230–250, 2013.

[11] R. Durrett. *Branching Process Models of Cancer*. Springer, 2014.

[12] R. Durrett et al. Evolutionary dynamics of tumor progression with random fitness values. *Journal of Theoretical Population Biology*, 78(1):54–66, 2010.

[13] R. Durrett and S. Moseley. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77(1):42–48, 2010.

[14] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.

[15] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[16] J. Foo and K. Leder. Dynamics of cancer recurrence. *The Annals of Applied Probability*, 23(4):1437–1468, 2013.

[17] P. A. Futreal. Intra-tumor heterogeneity in localized lung adenocarcinomas delineated by multi-region sequencing. *Science*, 346:256–259, 2014.

[18] H. Haeno and F. Michor. The evolution of tumor metastases. *Journal of Theoretical Biology*, (263):30–44, 2010.

[19] A. Hamon and B. Ycart. Statistics for the luria-delbrück distribution. *Electronic Journal of Statistics*, 6:1251–1272, 2012.

[20] Y. Iwasa, M. A. Nowak, and F. Michor. Evolution of resistance during clonal expansion. *Genetics*, 172(4):2557–2566, 2006.

[21] S. Janson. Functional limit theorems for multitype branching processes and generalized pólya urns. *Stochastic Processes and their Applications*, 110(2):177–245, 2004.

[22] S. Janson. Limit theorems for triangular urn schemes. *Probability Theory and Related Fields*, 134(3):417–452, 2006.

[23] S. Jones et al. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11):4283–4288, 2008.

[24] P. Keller and T. Antal. Mutant number distribution in an exponentially growing population. *Journal of Statistical Mechanics: Theory and Experiment*, P01011, 2015.

[25] D. G. Kendall. Birth-and-death processes, and the theory of carcinogenesis. *Biometrika*, 47(1-2):13–21, 1960.

[26] D. A. Kessler and H. Levine. Large population solution of the stochastic Luria-Delbrück evolution model. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29):11628–11687, 2013.

[27] D. A. Kessler and H. Levine. Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. *Journal of Statistical Physics*, 158(4):783–805, 2015.

[28] N. Komarova. Stochastic modeling of drug resistance in cancer. *Journal of Theoretical Biology*, 239(3):351–366, 2006.

[29] N. Komarova, L. Wu, and P Baldi. The fixed-size Luria-Delbrück model with a nonzero death rate. *Mathematical Biosciences*, 210(1):253–290, 2007.

[30] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome research*, 2017.

[31] D. E. Lea and C. A. Coulson. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics*, 49(3):264–285, 1949.

[32] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 48(6):419–511, 1943.

[33] F. Michor, M. A. Nowak, and Y. Iwasa. Stochastic dynamics of metastasis formation. *Journal of Theoretical Biology*, 240(4):521–530, 2006.

[34] M. D. Nicholson and T. Antal. Universal asymptotic clone size distribution for general population growth. *Bulletin of Mathematical Biology*, 78(11):2243–2276, 2016.

[35] H. Ohtsuki and H. Innan. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theoretical Population Biology*, 117:43–50, 2017.

[36] A. Rényi. On the theory of order statistics. *Acta Mathematica Hungarica*, 4(3-4):191–231, 1953.

[37] W. A. Rosche and P. L. Foster. Determining mutation rates in bacterial populations. *Methods*, 20(1):4–17, 2000.

[38] M. J. Williams et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50:895–903, 2018.

[39] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48:238–244, 2016.

[40] T. I. Zack et al. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45:1134–1140, 2013.

[41] Q. Zheng. Progress of a half century in the study of the Luria-Delbrück distribution. *Mathematical Biosciences*, 162(1-2):1–32, 1999.